Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Introduction to statistical inference
In the present module we are going to introduce the concept of point estimation which is a part of statistical inference. Statistical inference is the process of going from information gained from a sample to infer about a population from which the sample is taken. There are two aspects of statistical inference that we will be studying in this course: (i) Estimation and (ii) Hypothesis testing. In an estimation problem some features of the population in which an enquirer is interested may be completely unknown to him, and he may want to make a guess about this feature completely on the basis of a random sample from the population. There are two types of estimation problem : (i) Point estimation and (ii) Interval estimation. In this lecture we shall discuss some preliminary concept of point estimation.Let us start our discussion with a brief history of estimation problem.

Historical Perspective The problem of estimation arose in a very natural way in problems of Astronomy and Geodesy in the first half of the 18-th century. For example in Astronomy, the determination of interplanetary distances, determining the position of planets and their movements in time were some of the important problems. Whereas in Geodesy, determining the spheroidical shape of the earth was one of the most important problems. It is known that the figure of the earth is almost a sphere except for some flatness near the poles. Observations were obtained on the measurement of the length of one degree of a certain meridian and the problem was to determine the parameters say α and β which specified the spheroid of the earth. Indirect observations on (α, β) were given by the relation

$$Y_i = \alpha + \beta x_i, i = 1, 2, \ldots, n$$

Where x_i's are known fixed constants. Note that (α, β) are uniquely determined if only two observations on Y at different values of (x_1, x_2) are available.However, as is customary in science, several observations were made at different values $(x_1, x_2, \ldots x_n)$ and this led to the theory of combination of observations with random error which directly or indirectly measured "magnitudes of interest or parameters. To estimate α and β on the basis of the

given data the first attempt was made by Rogerr Boscovich(1757) in course of a geodetic study of ellipticity (extent of flatness at the poles) of the earth. He suggested that the estimates of (α, β) are to be determined such that
(i) The sum of positive and negative residuals or errors should balance i.e. $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$ and
(ii) Subject to the above constraint we determine (α, β) such that $R = \sum_{i=1}^{n}|y_i - \alpha - \beta x_i|$ The sum of absolute values of errors $e_i = y_i - \alpha - \beta x_i$ is as small as possible.
Using geometric argument Boscovich solve the problem for the five observations that he had. Laplace (1789) gave a general algebric algorithm to obtain estimates of (α and β) on the above principles for any number of observations. This problem was later solved by Gauss and Legendre using the method of least squares.
Boscovich has made the assumption that the errors of overestimation and underestimation must balance out. This idea was used by so many researchers in future time. For estimating the parameter θ in the simplest model $Y_i = \theta + e_i$, Simpson(1776) used this idea by assuming that errors are symmetrically uniformly distributed about zero or the probability density function of the error is given by $f(e) = \frac{1}{2h}, -h < e < h, h > 0$. Euler (1778) proposed the arc of a parabolic curve given by $f(e) = \frac{3}{4r^3}(r^2 - e^2), -r < e < r, r > 0$ as the pdf of the random error. Laplace suggested the probability density function $f(e) = \frac{1}{2h} exp\left[\frac{-|e|}{h}\right], -\infty < e < \infty$. As the model for distribution of errors and Gauss proposed the normal distribution with probability density function $f(e) = \frac{1}{\sqrt{2\pi h^2}} exp\left[\frac{-e^2}{2h^2}\right], -\infty < e < \infty$. It is important to point out here that the double exponential distribution used by Laplace to represent error distribution led to the median of the sample of the "best" estimator of the "True Value" of the parameter θ whereas the normal distribution used by Gauss led to the mean of the sample as the "best" estimator of the "True value".

Theory of Point estimation

Background

We consider a random experiment E. The outcome of E is represented by a Observable random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n), n \geq 1$. A particular value of $\boldsymbol{X}$ is denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. The character $\boldsymbol{X}$ could be real or vector valued and the set of all values of $\boldsymbol{X}$ is called the sample space and it is denoted by $\mathcal{X} \subset \mathcal{R}^n$.
The random vector $\boldsymbol{X}$ is generated by $F(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$, the distribution function of $\boldsymbol{X}$.

In a parametric point estimation problem we assume that the functional form of $F(\boldsymbol{x})$ is known except perhaps for a certain number of parameters. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ be the unknown parameters associated with $F(\boldsymbol{x})$. The parameter θ may be real valued or vector valued and is usually called a labelling or indexing parameter. The labelling parameter θ varies over a set of values, called as parameter space and is denoted by $\Theta \subset \mathcal{R}^k$. So $F(\boldsymbol{x})$ can be looked upon as a function of $\boldsymbol{\theta}$ and henceforth we will write it as $F_\theta(\boldsymbol{x})$. If $\boldsymbol{X}$ is discrete or absolutely continuous then $F(\boldsymbol{x})$ is generated by $f_\theta(\boldsymbol{x})$, the probability mass function(p.m.f.) or of probability density function (p.d.f.) $\boldsymbol{X}$. We write $\mathcal{F}_\theta = \{p(\boldsymbol{x}, \theta) : \boldsymbol{\theta} \in \Theta\}$, as the class of all probability mass or density functions. The object of inference is the parameter θ or a function of the parameter θ say $g(\theta)$, which is of interest. Let us consider few examples.

Example 1 Suppose a coin is tossed 50 times.
The outcome of ith toss can be described by a random variable X_i such that $X_i = 1$or 0 according as the ith toss results in a head or a tail.

$$\mathcal{X} = \{(x_1, x_2, \ldots x_{50}) : x_i = 0 \ or \ 1 \text{ for all } i\}$$

If θ be the probability of getting a head in any toss then $\Theta = (0, 1)$ and the probability function of $\boldsymbol{X}$ is $p(\boldsymbol{x}, \theta) = \prod_{i=1}^{50} \theta^{x_i}(1-\theta)^{1-x_i}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$. We may want to estimate θ or any function of θ.

Example 2 Suppose that 100 seeds of a certain flower were planted one in each pot and let X_i equal one or zero according as the seed in the ith pot germinates or not. The data consists of $(x_1, x_2, \ldots x_{100})$ a sequence of ones and zeroes and is regarded as a realization of $(X_1, X_2, \ldots X_{100})$ such that components are i.i.d. random variables with $P[X_1 = 1] = \theta$ and $P[X_1 = 0] = 1 - \theta$, where θ represents the probability that a seed germinates. The object of estimation is θ itself or a function $g(\theta)$ that may be of interest. For example, consider $g(\theta) = \binom{10}{8}\theta^8(1-\theta)^2$, which is the probability that in a batch of 10 seeds exactly 8 seeds will germinate.

Example 3 In a pathbreaking experiment Rutherford Chadwick and Elis(1920) observed 2608 time intervals of 7.5 seconds each and counted the number of time intervals N_r in which exactly r number of α particles hit the counter. They obtained the following table

r	0	1	2	3	4	5	6	7	8	9	≥ 10
N_r	57	203	383	525	532	408	273	139	45	27	18

It is quite well known that the Poisson distribution with p.m.f $f_\theta(x) = \frac{\exp(-\theta)\theta^x}{x!}, x = 1, 2, \ldots \theta > 0$ serves as a good model for the number of times a given event E occurs in a unit time interval. If X_i denotes the number of α particles hitting the counter in the $i-th$ time interval then $(X_1, X_2, \ldots X_n)$ where $n = 2608$ are i.i.d. Poisson random variables with parameter θ. We may want to estimate θ on the basis of the given data.

Example 4 Consider determination for an ideal physical constant such as gravity g. Usual way to estimate g is by the pendulum experiment and observe $X = \frac{4\pi^2 l}{T}$, where l is the length of the pendulum and T the time required for a fixed number of oscillations. Due to variation which depends on several factors such as the skill of the experimentor and measurement errors, the $i-th$ observation $X_i = g + e_i$ where e_i is the random error. Assuming the distribution of error is normal with zero mean and variance σ^2 we have $X_1, X_2, \ldots, X_n$ are i.i.d.$N(g, \sigma^2)$. Here the parameter θ is a two dimensional vector, $\theta = (g, \sigma^2)$. Here we can view estimation of g. On the other hand one may be interested in estimating the error variance σ^2 through which we can estimate the ability of the experimenter.

Example 5 Suppose an experiment is conducted by measuring the length of lives in hours of n electric bulbs produced by a certain company.
Let X_i be the length of live for the ith bulb.

$$\mathcal{X} = \{(x_1, x_2, \ldots x_n) : x_i \geq 0 \text{ for all } i\}$$

If we assume that the distribution of each X_i is exponential with mean θ then $\Theta = (0, \infty)$ and and the probability function of $\boldsymbol{X}$ is $p(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$. We may want to estimate the parameter θ or $g(\theta) = e^{-\frac{60}{\theta}}$, which represents the probability that the lifetime of a bulb will be at least 60 hours.

Objective
The distribution of $\boldsymbol{X}$ is characterized by the unknown parameter $\boldsymbol{\theta}$ about which we only know that it belongs to the parameter space Θ. To discuss the problem of point estimation, for the sake of simplicity, we consider the case when the parameter of interest is a real valued function $g = g(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$. In point estimation we try to approximate $g(\boldsymbol{\theta})$ on the basis of the observed value $\boldsymbol{x}$ of $\boldsymbol{X}$. In other words we try to put forward a particular statistic or a function of $\boldsymbol{X}$, say $T = T(\boldsymbol{X})$, which would represent the unknown $g(\boldsymbol{\theta})$

very closely. Such a statistic T is called an estimator or a point estimator of $g(\boldsymbol{\theta})$. Mathematically, T is a measurable mapping from $\mathcal{X}$ to the space of $g(\boldsymbol{\theta})$ and it is called an admissible estimator. Any observed value of T is called an estimate of $g(\boldsymbol{\theta})$. In a nutshell, a point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called a point estimator of θ. It is to be noted that for a particular estimator T for a parameter θ the estimate of θ may vary from sample to sample.

Example Suppose we want to estimate θ in Example 1.We may use the statistic $T = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of θ. Here T is a mapping from $\mathcal{X}$ to $(0, 1)$ and it is admissible.

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In any given problem of estimation we may have a large, often an infinite class of competing estimators for $g(\theta)$, a real valued function of the parameter θ.The following question that may arise is : Are some of many possible estimators better, in some sense, than others? In this section we will define certain criteria, which an estimator may or may not possess, that will help us in comparing the performances of rival estimators and deciding which one is perhaps the 'best'.

Closeness

If our object is to estimate a parametric function $g(\theta)$ then we would like the estimator $T(X)$ to be close to $g(\theta)$. Since $T(X)$ is a statistic, the usual measure of closeness $|T(X) - g(\theta)|$ is also a random variable and as a measure of closeness of T we use the measure $P_\theta(\mid T_1 - g \mid < \epsilon)$ for some $\epsilon > 0$.

Consider two estimators T_1 and T_2 for estimating a parametric function $g = g(\theta)$ of θ. The estimator T_1 will be called more concentrated estimator of $g(\theta)$ than T_2 if T_1 if for every $\epsilon > 0$

$$P_\theta(\mid T_1 - g \mid < \epsilon) \geq P_\theta(\mid T_2 - g \mid < \epsilon), \text{ for all } \ \theta \in \Theta \tag{1}$$

Result : A necessary condition for (1) to hold is that

$$E_\theta(T_1 - g)^2 \leq E_\theta(T_1 - g)^2, \text{ for all } \ \theta \in \Theta$$

provided $E_\theta(T_i - \theta)^2$ exists for all i=1,2.

Proof (For continuous case): We know that for every non-negative random variable X such that E(X) exists,

$$E(X) = \int_0^\infty P(X > x)dx.$$

Since $(T_1 - \theta)^2$ is a non-negative random variable we get

$$E_\theta(T_1 - g)^2 = \int_0^\infty P_\theta(\mid T_1 - g \mid > \epsilon) d\epsilon$$

It follows that

$$E_\theta(T_1 - g)^2 - E_\theta(T_2 - g)^2 = \int_0^\infty \left[P_\theta(\mid T_2 - g \mid < \epsilon) - P_\theta(\mid T_1 - g \mid < \epsilon)\right] d\epsilon$$

Hence the inequality $E_\theta(T_1 - g)^2 \leq E_\theta(T_2 - g)^2$ for all $\theta \in \Theta$ implies that $P_\theta(\mid T_2 - g \mid < \epsilon) \leq P_\theta(\mid T_1 - g \mid < \epsilon)$ for all $\epsilon > 0$.

Mean-squared Error(MSE)

If T be an estimator of g then the MSE of T is defined by

$$MSE_\theta(T) = E_\theta(T - g)^2, \text{ for all } \theta \in \Theta$$

The term $(T - g)$ is called the error of T in estimating g and thus $E(T - g)^2$ is called the mean square error of T. It measures the average squared difference between the estimator T and the parameter g. From the above result it is clear that smaller the value of MSE the better is the estimator. Naturally, we would prefer an estimator with smaller or smallest MSE. If such an estimator exists it will be best for the parameter g.

An estimator T is said to be best for g if $MSE_\theta(T) < MSE_\theta(T')$ for all $\theta \in \Theta$ for any other estimator T' of g. But the problem is that no such best estimator will exists in this sense. It will be clear from the following discussion.

Let for a particular value of θ, say θ_0, T' be defined as

$$T' = g(\theta_0) \text{ for all } \boldsymbol{x} \in \mathcal{X}$$

Then

$$\begin{aligned} MSE_{\theta_0}(T') &= E_{\theta_0}\left[g(\theta_0) - g(\theta_0)\right]^2 = 0 \\ \Rightarrow MSE_{\theta_0}(T) &= 0 \end{aligned}$$

Hence $T = g(\theta_0)$ with probability 1. Since θ_0 is arbitrary for any θ

$$T = g(\theta) \text{ with probability } 1$$

But T being a statistic it can not be a function of the unknown θ. Hence such a best estimator does not exist.

Consider the following example.

Example Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\theta, 1), \theta in \mathcal{R}$random variables. To estimate θ let us consider two estimators $T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T' = \theta_0$.

Then we get

$MSE_\theta(T) = \frac{1}{n}$ and $MSE_\theta(T') = (\theta_0 - \theta)^2$.

Now for values of $\theta \in [\theta_0 - \frac{1}{\sqrt{n}}, \theta_0 + \frac{1}{\sqrt{n}}]$ we have $MSE_\theta(T') \leq MSE_\theta(T)$ and for other values of θ we have $MSE_\theta(T') > MSE_\theta(T)$.

Here T' is not a good estimator of θ sine it always estimate θ to be θ_0 and it does not depend on observations at all. On the other hand the estimator T utilizes the observations and therefore it is better than T'.

From the above discussion it is clear that if MSE is the only criterion in search for a good estimator then there may be some 'freak' estimators (like T') that are extremely prejudiced in favour of a particular values of θ and they would perform better than a generally good estimator estimator at those points. For instance, in the above example the estimator T' is highly partial to θ_0 since it always estimate θ to be θ_0. One could restrict such freak estimators by considering only estimators that satisfy some other property. One such property is that of unbiasedness.

Unbiasedness

<u>Definition</u> An Estimator T is said to be an unbiased estimator (UE) of $g(\theta)$ if

$$E_\theta(T) = g(\theta) \text{ for all } \theta \in \Theta$$

If T is not an unbiased estimator of $g(\theta)$ then the bias of T is defined by

$$B_\theta(T) = E_\theta(T) - g(\theta), \theta \in \Theta.$$

The following result shows a relationship between MSE and variance of an estimator in terms of the bias.

Result $MSE_\theta(T) - Var_\theta(T) = B_\theta^2(T), \theta \in \Theta.$

Proof

$$\begin{aligned}
MSE_\theta(T) &= E_\theta\left(T - g(\theta)\right)^2 \\
&= E_\theta\left[(T - E_\theta(T)) + (E_\theta(T) - g(\theta))\right]^2 \\
&= E_\theta\left(T - E_\theta(T)\right)^2 + \left(E_\theta(T) - g(\theta)\right)^2 + 2\left(E_\theta(T) - g(\theta)\right) E_\theta\left(T - E_\theta(T)\right) \\
&= E_\theta\left(T - E_\theta(T)\right)^2 + \left(E_\theta(T) - g(\theta)\right)^2 \\
&= Var_\theta(T) + B_\theta^2(T)
\end{aligned}$$

Thus, MSE incorporates two components one measuring the variability of the estimator(precision) and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias. To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. Clearly, unbiased estimators do a good job of controlling bias. For an unbiased estimator T we have $MSE_\theta(T) = Var_\theta(T)$.

Although many unbiased estimators may be reasonable from the standpoint of MSE, but controlling bias does not guarantee that MSE is controlled. In some cases a trade-off occurs between the variance and the bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting a smaller MSE. It is clear from the following example.

Example 1 Let $X_i \sim N(\mu, \sigma^2) i = 1, 2 \ldots n$ independently where μ and σ^2 are unknown.

Consider all estimators of σ^2 of the form $T = cS^2$ where $c > 0$ is a constant and

$S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$

Now

$$MSE_\sigma(cS^2-\sigma^2)^2 = E_\sigma(cS^2-\sigma^2)^2 = c^2E_\sigma(S^4) - 2cE_\sigma(S^2) + \sigma^4$$

Since $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,

$$E_\sigma\left(\frac{(n-1)S^2}{\sigma^2}\right) = (n-1), V_\sigma\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$$

which gives $E_\sigma(S^2) = \sigma^2$ and $V_\sigma(S^2) = \frac{2\sigma^4}{n-1}$. After some routine algebra we get

$$MSE_\sigma(cS^2-\sigma^2)^2 = \sigma^4\left[c^2\frac{n+1}{n-1} - 2c + 1\right]$$

which attains the minimum when $c = \frac{n-1}{n+1}$. The minimum value being

$$\frac{2\sigma^4}{n+1} < \frac{2\sigma^4}{n-1}$$

and hence $T = \frac{1}{n+1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$ has smaller MSE than the unbiased estimator S^2 of σ^2. But T is not unbiased for σ^2 since $E_\sigma(T) = \frac{n-1}{n+1}\sigma^2$. If we use the criterion of MSE the estimator T which is a biased for σ^2 is better than the unbiased estimatorS^2 of σ^2.

Example 2 Let $X_i \sim N(\theta, 1) i = 1, 2 \ldots n$ independently where $|\theta| \leq 1$. Here $\bar{X}$ is unbiased for θ. Let us consider the following estimator T of θ such that

$$\begin{aligned} T &= -1 \text{ if } \bar{X} < -1 \\ &= \bar{X} \text{ if } \mid \bar{X} \mid \leq 1 \\ &= 1 \text{ if } \bar{X} > 1 \end{aligned}$$

Then

$$E_\theta(T) = P_\theta(\bar{X} > 1) + P_\theta(\bar{X} < -1) + \int_{-1}^{1} \bar{x} f(\bar{x}) d\bar{x} \neq \theta,$$

where $f(\bar{x})$ represents the p.d.f. of $\bar{X}$. Hence T is a biased estimator of θ and the MSE of T is given by

$$MSE_\theta(T) = (1-\theta)^2 P_\theta(\bar{X} > 1) + (-1-\theta)^2 P_\theta(\bar{X} < -1) + \int_{-1}^{1} (\bar{x}-\theta)^2 f(\bar{x}) d\bar{x}$$

For the estimator $\bar{X}$ we have

$$E_\theta(\bar{X}) = \theta.$$

Hence $\bar{X}$ is an unbiased estimator of θ and

$$MSE_\theta(\bar{X}) = \int_{-\infty}^{\infty} (\bar{x}-\theta)^2 f(\bar{x}) d\bar{x}.$$

Now

$$\begin{aligned} MSE_\theta(\bar{X}) - MSE_\theta(T) &= \int_{-\infty}^{\infty} (\bar{x}-\theta)^2 f(\bar{x}) d\bar{x} - MSE_\theta(T) \\ &= \int_{-\infty}^{-1} \left[(\bar{x}-\theta)^2 - (-1-\theta)^2\right] f(\bar{x}) d\bar{x} + \int_{1}^{\infty} \left[(\bar{x}-\theta)^2 - (1-\theta)^2\right] f(\bar{x}) d\bar{x} \\ &= I_1 + I_2, say. \end{aligned}$$

In the first integral I_1,

$$\bar{x} < -1 \Rightarrow (\bar{x}-\theta) < (-1-\theta) < 0 \Rightarrow (\bar{x}-\theta)^2 > (-1-\theta)^2.$$

In the second integral I_2,

$$\bar{x} > 1 \Rightarrow (\bar{x}-\theta) > (-1-\theta) > 0 \Rightarrow (\bar{x}-\theta)^2 > (-1-\theta)^2.$$

Hence we get $I_1 + I_2 > 0$. Thus T is a biased estimator of θ but the MSE of T is less than that of $\bar{X}$.

Note In both the examples a natural question may arise : which then should be preferred? The answer obviously depends on the purpose for which an estimate is obtained.

e-pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Estimable Function: A parametric function $g(\theta)$ is called estimable if there exists an atleast one unbiased estimator of $g(\theta)$.

Result: If $X \sim Bin(n, \theta)$ the estimable functions of θ are polynomial functions of θ of degree n or less.

Proof: Let $g(\theta)$ be an estimable function of θ. Then there exists a statistic $T(X)$ such that

$$E_\theta(T(X)) = g(\theta) \text{ for all } \theta \in (0,1)$$

$$\Rightarrow \sum_{i=0}^{n} T(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = g(\theta) \text{ for all } \theta \in (0,1) \tag{2}$$

Since the L.H.S. of (2) is a polynomial in θ of degree n or less, to hold the identity the R.H.S. must be a polynomial in θ of degree n or less.

Note: If $X \sim Poisson(\theta)$, the estimable functions of θ are convergent power series in θ.

Example 1: Consider a random variable X which takes only two values 1 or 2 with respective probabilities $\theta^2 + \theta^3$ and $1 - \theta^2 - \theta^3$. Here θ is an unknown parameter whose only possible values are 1/3 and 2/3. On the basis of a single observation X, we want to see whether the function $\frac{1}{\theta}$ is estimable or not.

If possible suppose $\frac{1}{\theta}$ is estimable. Then there exists a statistic T(X) which is an unbiased estimator of $\frac{1}{\theta}$. Hence

$$E_\theta(T) = \frac{1}{\theta}, \text{ for } \theta = 1/3, 2/3$$

$$\Rightarrow T(1)(\theta^2 + \theta^3) + T(2)(1 - \theta^2 - \theta^3) = \frac{1}{\theta}, \text{ for } \theta = 1/3, 2/3$$

We get

$$\frac{4}{27}T(1) + \frac{23}{27}T(2) = 3$$

$$\frac{20}{27}T(1) + \frac{7}{27}T(2) = 3/2$$

Two equations are satisfied when $T(1) = \frac{27}{32}$ and $T(2) = \frac{27}{8}$. Hence there exists an unbiased estimator of $\frac{1}{\theta}$ that is $\frac{1}{\theta}$ is estimable.

Example 2 : Let $X_i \sim N(\mu, \sigma^2) i = 1, 2 \ldots n$ independently where μ and σ^2 are unknown.

Let $\theta = (\mu, \sigma^2)$ and $g(\theta) = e^{-\mu + \frac{1}{2}\sigma^2}$.

From the m.g.f. of the normal distribution

$E_\theta(e^{tX}) = e^{t\mu + \frac{1}{2}t^2\sigma^2}$ for all θ and t

Setting $t = -1$ we get $E_\theta(e^{-X}) = e^{-\mu + \frac{1}{2}\sigma^2}$ for all θ and hence

$$E_\theta\left(\frac{1}{n}\sum_{i=1}^{n} e^{-X_i}\right) = e^{-\mu + \frac{1}{2}\sigma^2} \text{ for all } \theta.$$

Thus there exists an unbiased estimator of $g(\theta)$ and hence $g(\theta)$ is estimable.

Best linear Unbiased Estimator (BLUE)

Let $T_1, T_2, \ldots T_k$ be k unbiased estimators of $g(\theta)$. Consider a linear combination of $T_1, T_2, \ldots T_k$,namely, $T = \sum_{i=1}^{k} l_i T_i$ where $l_1, l_2 \ldots l_k$ are real numbers. For T to be unbiased we must have

$$E_\theta(T) = g(\theta) \text{ for all } \theta$$

$$\Rightarrow \sum_{i=1}^{k} l_i E_\theta(T_i) = g(\theta) \text{ for all } \theta$$

$$\Rightarrow g(\theta)\sum_{i=1}^{k} l_i = g(\theta) \text{ for all } \theta$$

$$\Rightarrow \sum_{i=1}^{k} l_i = 1, \text{by equating the coefficients of } g(\theta)$$

Such an unbiased estimator T is called a linear unbiased estimator of $g(\theta)$. The estimator T will be called the BLUE for $g(\theta)$ if it has the minimum variance among all linear unbiased estimators of $g(\theta)$.

Determination of BLUE : For the sake of simplicity we assume that T_i's are independent. Let $Var_\theta(T_i) = \sigma_i^2$ for all $i = 1, 2, \ldots k$. Then the variance of T is given by

$$Var_\theta(T) = \sum_{i=1}^{k} l_i^2 Var_\theta(T_i) = \sum_{i=1}^{k} l_i^2 \sigma_i^2.$$

Our problem is to minimize $\sum_{i=1}^{k} l_i^2\sigma_i^2$ subject to the constraint $\sum_{i=1}^{k} l_i = 1$.

By Cauchy-Schwarz inequality, for any two sets of real numbers $a_1, a_2, \dots a_k$ and $b_1, b_2, \dots b_k$

$$\left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right) \geq \left(\sum_{i=1}^{k} a_i b_i\right)^2,$$

equality sign holds if $a_i = cb_i$ for all $i = 1, 2, \dots k$, where c is a constant.

Setting $a_i = l_i\sigma_i$ and $b_i = \frac{1}{\sigma_i}$ for all $i = 1, 2, \dots k$ we get

$$\left(\sum_{i=1}^{k} l_i^2\sigma_i^2\right)\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right) \geq \left(\sum_{i=1}^{k} l_i\right)^2$$

$$\Rightarrow \left(\sum_{i=1}^{k} l_i^2\sigma_i^2\right) \geq \frac{1}{\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)}$$

The minimum value of $Var_\theta(T)$ is $\frac{1}{\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)}$ and the minimum is attained if $l_i\sigma_i^2 = c$ for all i implying that $l_i = \frac{c}{\sigma_i^2}$ for all i. Using the constraint $\sum_{i=1}^{k} l_i = 1$ we get $c = \frac{1}{\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)}$ and hence $l_i = \frac{w_i}{\sum_{i=1}^{k} w_i}$ where $w_i = \frac{1}{\sigma_i^2}$ for all i. The BLUE for $g(\theta)$ is given by $\frac{\sum_{i=1}^{k} T_i w_i}{\sum_{i=1}^{k} w_i}$ which is the weighted mean of T_i's with w_i's as weights.

Note : If $Var_\theta(T_i) = \sigma^2$ for all $i = 1, 2, \dots k$ then $T = \frac{\sum_{i=1}^{k} T_i}{k}$ which is the simple arithmatic mean of T_i's.

Some remarks on unbiased estimator:

Remark:1.An Unbiased Estimator may be inadmissable.

Example: Suppose $X \sim P(\theta), g(\theta) = e^{-3\theta}, \theta > 0$

We want to find an unbiased estimator of $g(\theta)$.

Let $T(X)$ be an unbiased estimator of $g(\theta)$. Hence

$$E_\theta\left(T(X)\right) = g(\theta) \quad \text{for all} \quad \theta > 0$$

$$\Rightarrow \sum_{x=0}^{\infty} T(x)\frac{e^{-\theta}\theta^x}{x!} = e^{-3\theta} \quad \text{for all} \quad \theta > 0$$

$$\Rightarrow \sum_{x=0}^{\infty} T(x)\frac{\theta^x}{x!} = \sum_{x=0}^{\infty}(-2)^x\frac{\theta^x}{x!} \quad \text{for all} \quad \theta > 0$$

Comparing the coefficients of $\frac{\theta^x}{x!}$ from both sides of the above identity we get $T(x) = (-2)^x$ for all x. Hence $T(X) = (-2)^X$ is the unique unbiased estimator of $g(\theta)$ but it is improper since $g(\theta)$ is a monotone decreasing continuous function of θ taking only non-negative values whereas T is a an oscillatory function of X taking both positive and negative values.

Remark:2. An unbiased estimator may not exists.

Example: Let X be a random variable with p.m.f.

$$f_\theta(x) = (1-\theta)^x\theta, x = 0, 1, \ldots \infty, 0 < \theta < 1$$

Suppose our object is to find an unbiased estimator of θ. If possible suppose $T(X)$ be an unbiased estimator of θ.Then

$$E_\theta\left(T(X)\right) = \theta \quad \text{for all} \quad 0 < \theta < 1$$

$$\Rightarrow \sum_{x=0}^{\infty} T(x)(1-\theta)^x\theta = \theta \quad \text{for all} \quad 0 < \theta < 1$$

$$\Rightarrow \sum_{x=0}^{\infty} T(x)(1-\theta)^x = 1 \quad \text{for all} \quad 0 < \theta < 1$$

Such an identity can not hold since the L.H.S is a function of θ but the R.H.S is independent of θ.

Remark:3. An unbiased estimator may not be unique.

Example: Let $X_i \sim Rect(0,\theta), i = 1, 2 \ldots n$ independently. If $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $X_{(n)} = Max(X_1, X_2, \ldots X_n)$ then $T_1 = 2\bar{X}$ and $T_2 = \frac{n+1}{n}X_{(n)}$ are unbiased estimators of θ.

Proof: Since $X_i \sim Rect(0,\theta), i = 1, 2 \ldots n$,

$E_\theta(X_i) = \frac{\theta}{2}$ for all $\theta > 0$ and for all $i = 1, 2, \ldots n$.

It follows that $E_\theta(\bar{X}) = \frac{\theta}{2}$ for all $\theta > 0$ and hence $E_\theta(T_1) = \theta$ for all $\theta > 0$.

The p.d.f. of $X_{(n)}$ is $f_\theta(x_{(n)}) = \frac{n}{\theta^n}x_{(n)}^{n-1}, 0 < x_{(n)} < 1$

Hence $E_\theta(X_{(n)}) = \frac{n}{n+1}\theta$ for all $\theta > 0$ and hence $E_\theta(T_2) = \theta$ for all $\theta > 0$.

Remark 4. An unbiased estimator is not invariant under a transformation i.e.if T is an unbiased estimator of θ then $g(T)$ is not unbiased for $g(\theta)$.

Example : If g is convex downward then by Jensen's inequality $E_\theta(g(T)) \geq g(E(T))$ and hence $E_\theta(g(T)) \geq g(\theta)$ for all θ. For example suppose T is an unbiased estimator of θ and $g(\theta) = \theta^2$. Since g is convex downward we have $E_\theta(g(T)) \geq \theta^2$ for all θ i.e, T has an upward bias. Note that g is convex upward then $E_\theta(g(T)) \leq g(\theta)$ for all θ. For example suppose $g(\theta) = \frac{1}{\theta}, \theta \neq 0$. Since g is convex downward $E_\theta(g(T)) \leq \frac{1}{\theta}$ for all θ i.e, T has a downward bias.

A practical example:

1. Consider a coin with probability of getting a head in a toss is equal to $\theta, 0 < \theta < 1$, is unknown. To estimate $\frac{1}{\theta}$ unbiasedly the following inverse sampling methodology is adopted: The coin is tossed repeatedly until a head appears and the the experiment is performed independently for 20 times. From the outcomes of 20 trials it is found that the first head appears in 1st, 3rd, 5th, 1st, 2nd, 1st, 3rd, 7th, 2nd, 4th, 4th, 8th, 1st, 3rd, 6th, 5th, 2nd, 1st, 6th and 2nd toss.

Let X_i be the number of trials required to get the first head in the ith trial,i=1,2, $\ldots$20. The the p.m.f. of X_i is given by

$$f_\theta(x_i) = (1-\theta)^{x_i-1}\theta, x_i = 1, 2, \ldots; 0 < \theta < 1$$

Then

$$\begin{aligned} E_\theta(X_i) &= \sum_{x=1}^{\infty} x(1-\theta)^{x-1}\theta \\ &= \theta(1-(1-\theta))^{-2} \\ &= \frac{1}{\theta}, \quad \text{for all} \quad i \end{aligned}$$

Hence $E_\theta\left(\frac{1}{20}\sum_{i=1}^{20} X_i\right) = \frac{1}{\theta}$, for all θ From the give data we have $\sum_{i=1}^{20} x_i = 67$ and hence an unbiased estimate for $\frac{1}{\theta}$ is given by 3.35.

Note: From Remark 3 it is clear that it is not possible to find an unbiased estimate of θ from the above data set.

Statistical Inference I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Let $g(\theta)$ be an estimable function of the unknown parameter θ and $U_g = \{T(\boldsymbol{X}) : E_\theta(T) = g(\theta), 0 < V_\theta(T) < \infty, \theta \in \theta\}$ be the class of all unbiased estimators of $g(\theta)$ with finite variances. An estimator $T \in U_g$ is said to be a uniformly minimum variance unbiased estimator of $g(\theta)$ if $Var_\theta(T) \leq Var_\theta(T')$ for all $\theta \in \Theta, T' \in U_g$.

Method Of Covariance

Let $U_0 = \{h(\boldsymbol{X}) : E_\theta(h) = 0, 0 < V_\theta(h) < \infty, \theta \in \boldsymbol{\theta}\}$ be the class of all unbiased estimators of $g(\theta)$ with finite variances. Then an $T \in U_g$ is the UMVUE of g if and only if $Cov_\theta(T, h) = E_\theta(Th) = 0$ for all $\theta \in \Theta$ and for all $h \in U_0$ i.e, T is uncorrelated with every unbiased estimator of 0.

Proof : Only if part

Let $T \in U_g$ be the UMVUE of $g(\theta)$ and $T^\star$ be any other statistic such that

$T^\star = T + \lambda h$ where $h \in U_0$ and λ is a fixed constant.

Therefore $E_\theta(T^\star) = g(\theta)$ for all $\theta \in \Theta$

$\Rightarrow T^\star \in U_g$

Since T is the UMVUE

$$Var_\theta(T) \leq Var_\theta(T^\star) = Var_\theta(T) + \lambda^2 Var_\theta(h) + 2\lambda Cov_\theta(T, h) \text{ for all } \theta$$

$$\Rightarrow \lambda^2 Var_\theta(h) + 2\lambda Cov_\theta(T, h) \geq 0 \text{ for all } \theta \tag{1}$$

Now if $Cov_\theta(T, h) \neq 0$ for some $\theta_0 \in \Theta$ let us choose $\lambda = \frac{-Cov_{\theta_o}(T,h)}{Var_{\theta_o}(h)}$. Then for $\theta = \theta_0$ the L.H.S of (1) becomes

$$\lambda^2 Var_{\theta_o}(h) + 2\lambda Cov_{\theta_0}(T, h) = \frac{-Cov^2_{\theta_o}(T, h)}{Var_{\theta_o}(h)} < 0$$

This contradicts (1) for $\theta = \theta_0$.

Hence $Cov_\theta(T, h) = 0$ for all $\theta \in \Theta$.

If part Suppose $T \in U_g$ be an estimator of $g(\theta)$ such that $Cov_\theta(T, h) = 0$ for all $\theta \in \Theta$ and for all $h \in U_0$ Let $T^\star$ be another unbiased estimator of $g(\theta)$ and we define

$$h^\star = T^\star - T$$

Then $E_\theta(h^\star) = E_\theta(T^\star) - E_\theta(T) = 0$ for all $\theta \in \Theta$

$$\Rightarrow h^\star \in U_0$$

$$\Rightarrow Cov_\theta(T, h^\star) = 0 \text{ for all } \theta \in \Theta$$

$$\Rightarrow Cov_\theta(T, T^\star - T) = 0 \text{ for all } \theta \in \Theta$$

Now $var_\theta(T^\star) = Var_\theta(T^\star - T + T) = var_\theta(T^\star - T) + var_\theta(T)$

$$\geq var_\theta(T^)\text{for all } \theta \in \Theta$$

Since $T^\star$ is arbitrary, T has the minimum variance among all unbiased estimators of $g(\theta)$ i.e. T is the UMVUE of $g(\theta)$.

Example 1: Let $X_1, X_2, \ldots X_n$ be a i.i.d. $N(\theta, 1)$ random variables where $-\infty < \theta < \infty$ is unknown. Suppose we want to find the UMVUE of θ. The joint p.d.f of $X_1, X_2, \ldots X_n$ is given by

$$f_\theta(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} exp\left[\frac{-1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right], \boldsymbol{x} \in \mathcal{R}^n$$

Let $h(\boldsymbol{X})$ be an unbiased estimator of 0 i.e.

$$E_\theta(h(\boldsymbol{X})) = 0 \text{ for all } \theta$$

$$\Rightarrow \int h(\boldsymbol{x}) f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 0 \text{ for all } \theta$$

Differentiating w.r.t. θ we get

$$\int h(\boldsymbol{x}) \sum_{i=1}^{n} (X_i - \theta) f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 0 \text{ for all } \theta$$

$$\Rightarrow \int h(\boldsymbol{x})\,(\bar{x}-\theta)\,f_\theta(\boldsymbol{x})d\boldsymbol{x} = 0 \text{ for all } \theta$$

$$\Rightarrow \int h(\boldsymbol{x})\bar{x} f_\theta(\boldsymbol{x})d\boldsymbol{x} = 0 \text{ for all } \theta, \text{ since } E_\theta(h(\boldsymbol{X})) = 0 \text{ for all } \theta$$

$$\Rightarrow E_\theta\left(h(\boldsymbol{X})\bar{X}\right) = 0 \text{ for all } \theta$$

Also $E_\theta(\bar{X}) = \theta$ for all θ. Hence $\bar{X}$ is the unbiased estimator of θ which is uncorrelated with every unbiased estimator of 0. By the above result $\bar{x}$ is the UMVUE of θ.

Example 2 : Let X be a random variable with p.m.f

$$\begin{aligned} f_\theta(x) &= \theta, \text{ if } x = -1 \\ &= (1-\theta)^2\theta^x, \text{ if } x = 0,1,2,\ldots \end{aligned}$$

Suppose we want to find the parametric functions $g(\theta)$ which admits of UMVUE and also to find such estimators.

Let $h(X)$ be an unbiased estimator of 0 i.e.

$$E_\theta(h(X)) = 0 \quad \text{for all } \theta$$

$$\Rightarrow \theta h(-1) + \sum_{x=0}^{\infty} h(x)(1-\theta)^2\theta^x = 0 \quad \text{for all } \theta$$

$$\Rightarrow \sum_{x=0}^{\infty} h(x)\theta^x = \frac{-\theta h(-1)}{(1-\theta)^2} = -\sum_{x=0}^{\infty} h(-1)x\theta^x \quad \text{for all } \theta$$

Comparing the coefficients of $\theta^x, x = 0,1,2,\ldots$ from both sides of the above identity we get

$$h(x) = -xh(-1), x = 1,2,\ldots \text{ and } h(0) = 0 \tag{1}$$

Let T be the UMVUE of $g(\theta)$

$$E_\theta(Th) = 0 \text{ for all } \theta$$

Hence by a similar argument we get

$$t(x)h(x) = -xt(-1)h(-1), x = 1,2,\ldots \text{ and } t(0)h(0) = 0 \tag{2}$$

Comparing (1) and (2) we get

$t(x) = t(-1), x = 1, 2, \ldots$ and $t(0)$ is arbitrary.

$T(X)$ is thus defined is the UMVUE of its expectation.

$$E_\theta(T) = g(\theta) \text{ for all } \theta$$

$$\Rightarrow g(\theta) = (1-\theta)^2 t(0) + t(-1)\{1-(1-\theta)^2\}$$

$$\Rightarrow g(\theta) = c_1(1-\theta)^2 + c_0 \text{ where } c_0 = t(-1) \text{ and } c_1 = t(0) - t(-1)$$

Hence an estimable function $g(\theta)$ will have the UMVUE if and if it is of the form $g(\theta) = c_1(1-\theta)^2 + c_0$ and the UMVUE is given by

$$\begin{aligned} T(X) &= c_0 + c_1 \text{ if } X = 0 \\ &= c_0 \text{ if } X = -1, 1, 2, \ldots \end{aligned}$$

Note : If we take $g(\theta) = \theta$, then it can not be put in the above form and hence there does not exist UMVUE of θ though there exist an unbiased estimator of θ which is given by

$$\begin{aligned} T(X) &= 1 \text{ if } X = -1 \\ &= 0 \text{ if } X = 0, 1, \ldots. \end{aligned}$$

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Some Results Uniformly Minimum Variance Unbiased Estimator

Result 1 : A uniformly minimum variance unbiased estimator, if exists, is unique.

Proof : If possible, suppose T_1 and T_2 be two UMVUE of $g(\theta)$.

Then $E_\theta(T_1) = E_\theta(T_2)$ and $Var_\theta(T_1) = Var_\theta(T_2)$ for all θ Deine a statistic T such that

$$T = \frac{1}{2}(T_1 + T_2)$$

Then $E_\theta(T) = g(\theta)$ for all θ and

$$Var_\theta(T) = \frac{1}{4}[Var_\theta(T_1) + Var_\theta(T_2) + 2Cov_\theta(T_1, T_2)] \tag{1}$$

If $\rho_\theta(T_1, T_2)$ be the correlation coefficient between T_1 and T_2 then
$Cov_\theta(T_1, T_2) = \rho_\theta(T_1, T_2)\sqrt{Var_\theta(T_1)Var_\theta(T_2)} = \rho_\theta(T_1, T_2)Var_\theta(T_1)$
From (1) we get

$$Var_\theta(T) = \frac{1}{2}[Var_\theta(T_1) + \rho_\theta(T_1, T_2)Var_\theta(T_1)$$

Since T_1 is an UMVUE of $g(\theta)$

$$Var_\theta(T_1) \leq Var_\theta(T) \text{ for all } \theta$$

$$\Rightarrow \rho_\theta(T_1, T_2) \geq 1 \text{ for all } \theta$$

$$\Rightarrow \rho_\theta(T_1, T_2) = 1 \text{ for all } \theta$$

$$\Rightarrow P_\theta\left[T_1 = a(\theta) + b(\theta)T_2\right] = 1 \text{ for all } \theta, \text{ for some constants } a(\theta), b(\theta)$$

$$\Rightarrow E_\theta(T_1) = a(\theta) + b(\theta)E_\theta(T_2) \text{ for all } \theta$$

$$\Rightarrow g(\theta) = a(\theta) + b(\theta)g(\theta) \text{ for all } \theta$$

From the above identity we get

$a(\theta) = 0$ and $b(\theta) = 1$ for all θ. Hence $P_\theta[T_1 = T_2] = 1$ for all θ,

Result 2 : The correlation coefficient between the UMVUE and any other unbiased estimator of $g(\theta)$ is always non-negative.

Proof: Let T be the UMVUE of $g(\theta)$ any T' be any other unbiased estimator of $g(\theta)$. Then

$$E_\theta(T) = E_\theta(T') = g(\theta) \text{ for all } \theta$$

Let us define a statistic $h = T - T'$.

Then $E_\theta(h) = 0$ for all θ and hence h is an unbiased estimator of 0 i.e. $h \in U_0$.

By the method of covariance

$$Cov_\theta(T, h) = 0 \text{ for all } \theta$$

$$\Rightarrow Cov_\theta(T, T - T') = 0 \text{ for all } \theta$$

$$\Rightarrow Cov_\theta(T, T') = Var_\theta(T) \text{ for all } \theta$$

The correlation coefficient between T and T' is given by

$$\rho_\theta = \frac{Cov_\theta(T, T')}{\sqrt{Var_\theta(T)Var_\theta(T')}} = \sqrt{\frac{Var_\theta(T)}{Var_\theta(T')}} \text{ for all } \theta$$

which is always non-negative.

Note : The efficiency of T' w.r.t. to T is defined by

$$e_\theta = \frac{Var_\theta(T)}{Var_\theta(T')} \text{ for all } \theta$$

From the above result we get

$$\rho_\theta = \sqrt{e_\theta} \text{ for all } \theta.$$

Result 3 : If $T_1, T_2 \ldots T_k$ are the UMVUE's of $g_i(\theta), i = 1, 2, \ldots k$, then $\sum_{i=1}^{k} a_i T_i$ is the UUMVUE of $\sum_{i=1}^{k} a_i g_i(\theta)$.

Proof : Since $E_\theta(T_i) = g_i(\theta)$ for all i, we have

$$E_\theta(\sum_{i=1}^{k} a_i T_i) = \sum_{i=1}^{k} a_i g_i(\theta) \text{ for all } \theta.$$

Hence $\sum_{i=1}^{k} a_i T_i$ is an unbiased estimator of $\sum_{i=1}^{k} a_i g_i(\theta)$. Since T_i is the UMVUE of $g_i(\theta)$ for all i,by the method of covariance

$$E_\theta(T_i h) = 0 \text{ for all } \theta \text{ and for all } h \in U_0$$

$$\Rightarrow E_\theta(\sum_{i=1}^{k} a_i T_i.h)0 \text{ for all } \theta \text{ and for all } h \in U_0$$

Hence $\sum_{i=1}^{k} a_i T_i$ is the UUMVUE of $\sum_{i=1}^{k} a_i g_i(\theta)$.

Result 4 : If $T_1, T_2 \ldots T_k$ are the UMVUE's of $g_i(\theta), i = 1, 2, \ldots k$, then $\prod_{i=1}^{k} T_i$ is the UMVUE of its expectation.

Proof: Let $h \in U_0$. Since T_1 is the UMVUE of $g_1(\theta)$ by the method of covariance

$$E_\theta(T_1 h) = 0 \text{ for all } \theta$$

$$\Rightarrow T_1 h \in U_0$$

Again T_2 is the UMVUE of $g_1(\theta)$ we get

$$E_\theta(T_2 T_1 h) = 0 \text{ for all } \theta$$

$$\Rightarrow T_2 T_1 h \in U_0$$

Hence $E_\theta(T_3 T_2 T_1 h) = 0$ for all θ. Proceeding in this way we finally get

$$E_\theta(\prod_{i=1}^{k} T_i h) = 0 \text{ for all } \theta.$$

Hence $\prod_{i=1}^{k} T_i$ is the UMVUE of its expectation.

Note: If T is the UMVUE of $g(\theta)$ then by the above result T^k is the UMVUE of its expectation where k is any positive integer.

Result 5 : If the sample consists of n independent observations from the same distribution, then the UMVUE's are symmetric in the observations.

Proof : Let $X_1, X_2, \ldots X_n$ be a random sample of size n drawn from a population with probability function $f_\theta(x)$ where θ is an unknown parameter. Suppose $T(X_1, X_2, \ldots X_n)$ be an unbiased estimator of $g(\theta)$ with $Var_\theta(T(X_1, X_2, \ldots X_n)) = \sigma^2$.Let us consider a symmetric function of $(X_1, X_2, \ldots X_n)$ viz.

$T^\star = \frac{1}{n!}\Sigma(T(X_{i_1}, X_{i_2}, \ldots X_{i_n}))$ where the summation being taken over all possible permutations of $(i_1, i_2, \ldots i_n)$ of $(1, 2, \ldots n)$. Then $E_\theta(T^\star) = g(\theta)$ for all θ and $Var_\theta(T^\star)$ is equal to

$$\frac{1}{n!^2}\Sigma Var_\theta(T(X_{i_1}, X_{i_2}, \ldots X_{i_n}))) + \frac{1}{n!^2}\Sigma Cov_\theta(T(X_{i_1}, X_{i_2}, \ldots X_{i_n}), T(X_{j_1}, X_{j_2}, \ldots X_{j_n}))) \tag{1}$$

Now by symmetry $Var_\theta(T(X_{i_1}, X_{i_2}, \ldots X_{i_n}) = \sigma^2$ for all θ, for any permutation of $(i_1, i_2, \ldots i_n)$. Also by Cauchy-Schwarz inequality

$$Cov_\theta(T(X_{i_1}, X_{i_2}, \ldots X_{i_n}), T(X_{j_1}, X_{j_2}, \ldots X_{j_n}) \leq \sigma^2 \text{ for all } \theta$$

for any two differnent permutations $(i_1, i_2, \ldots i_n)$ and $(j_1, j_2, \ldots j_n)$.Hence from (1) we get

$$Var_\theta(T^\star) \leq \sigma^2 \frac{n! + n!(n! - 1)}{n!^2} = \sigma^2 \text{ for all } \theta$$

Hence $T^\star$ is the UMVUE of $g(\theta)$.

Result 6: Let T_1 and T_2 be two unbiased estimators of $g(\theta)$ with efficiencies e_1 and e_2 respectively then

$$|\rho - \sqrt{e_1 e_2}| \leq \sqrt{(1 - e_1)(1 - e_2)}$$

where ρ is the correlation coefficient between T_1 and T_2.

Proof : Let T be the UMVUE of $g(\theta)$ with $Var_\theta(T) = \sigma^2$. Then $Var_\theta(T_i) = \frac{\sigma^2}{e_i}$ for all $i = 1, 2$. Let us define a statistic $T^\star$ such that

$$T^\star = \alpha T_1 + (1 - \alpha)T_2, \text{ for any constant } \alpha \in (0, 1)$$

Then

$$E_\theta(T^\star) = \alpha g(\theta) + (1 - \alpha)g(\theta) = g(\theta) \text{ for all } \theta$$

Hence $T^\star$ is also an unbiased estimator of $g(\theta)$ with

$$Var_\theta(T^\star) = \alpha^2 Var_\theta(T_1) + (1-\alpha)^2 Var_\theta(T_2) + 2\alpha(1-\alpha)Cov_\theta(T_1, T_2)$$

$$= \alpha^2 Var_\theta(T_1) + (1-\alpha)^2 Var_\theta(T_2) + 2\alpha(1-\alpha)\rho\sqrt{Var_\theta(T_1)Var_\theta(T_2)}$$

$$= \sigma^2\left[\alpha^2\frac{1}{e_1} + (1-\alpha)^2\frac{1}{e_2} + 2\alpha(1-\alpha)\rho\sqrt{\frac{1}{e_1}\frac{1}{e_2}}\right]$$

Since T is the UMVUE of $g(\theta)$

$$Var_\theta(T) \leq Var_\theta(T^\star) \text{ for all } \theta$$

$$\Rightarrow \alpha^2\frac{1}{e_1} + (1-\alpha)^2\frac{1}{e_2} + 2\alpha(1-\alpha)\rho\sqrt{\frac{1}{e_1}\frac{1}{e_2}} \geq 1$$

After some routine algebra we get

$$a\alpha^2 + b\alpha + c \geq 0, \tag{1}$$

where $a = \left(\frac{1}{e_1} + \frac{1}{e_2} - \frac{2\rho}{\sqrt{e_1 e_2}}\right), b = 2\left(\frac{\rho}{\sqrt{e_1 e_2}} - \frac{1}{e_2}\right), c = \left(\frac{1}{e_2} - 1\right).$

From (1) it is clear that the discriminant of the quadratic equation $a\alpha^2 + b\alpha + c = 0$ can not be positive, we have

$$\left(\frac{\rho}{\sqrt{e_1 e_2}} - \frac{1}{e_2}\right)^2 \leq \left(\frac{1}{e_1} + \frac{1}{e_2} - \frac{2\rho}{\sqrt{e_1 e_2}}\right)\left(\frac{1}{e_2} - 1\right)$$

$$\Rightarrow \left(\frac{\rho}{\sqrt{e_1 e_2}} - 1\right)^2 \leq \left(1 - \frac{1}{e_1}\right)\left(1 - \frac{1}{e_2}\right)$$

$$\Rightarrow |\rho - \sqrt{e_1 e_2}| \leq \sqrt{(1-e_1)(1-e_2)}.$$

Note : Let T be the UMVUE and T' be any other estimator of $g(\theta)$ with efficiency e. If ρ is the correlation coefficient between T and T' then from the above result we have

$$1 - 2e \leq \rho \leq 1.$$

Statistical Inference I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

SUFFICIENCY

Let $X_1, X_2, \ldots X_n$ be a random sample of size n from a population with p.m.f./p.d.f. $f_\theta(x)$, where $\theta \in \Theta \subset \mathcal{R}$ is an unknown parameter. A sufficient statistic for θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$. In other words, if a statistic $T(\boldsymbol{X})$ is sufficient for θ then after observing T it is not possible to get any additional information about θ from the sample.

Definition A statistic T is said to be an sufficient statistic for θ (or, rather, for the family of distributions $\mathcal{F}_\theta = \{f_\theta(\mathbf{x}) : \theta \in \Theta\}$) if the conditional distribution of any other statistic, say $T^\star$, given $T = t$ is independent of θ for every possible values of t.

Note : If we take $T^\star = \boldsymbol{X} = (X_1, X_2, \ldots X_n)$, then a statistic T is said to be a sufficient statistic for θ if, the conditional distribution of $(X_1, X_2, \ldots X_n)$ given $T = t$ is independent of θ for every possible values of t.

Remark 1 In this definition the parameter θ and the statistic T may be vector valued.

Remark 2 If T is sufficient for the family $\mathcal{F}_\theta = \{f_\theta(\mathbf{x}) : \theta \in \Theta\}$ then T is sufficient for the family $\mathcal{F}^\star{}_\theta = \{f_\theta(\mathbf{x}) : \theta \in \Theta^\star\}$ where $\Theta^\star \subset \Theta$.

Remark 3 A sufficient statistic contains all information regarding θ in the sense that from the knowledge of T alone it is possible to generate (by a random mechanism) a random quantity $\boldsymbol{Y}$ which is completely equivalent to the original $\boldsymbol{X}$ i.e. the distribution of $\boldsymbol{Y}$ is same as that of $\boldsymbol{X}$. Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution for all θ, they provide exactly same information about θ.

To justify this statement, suppose $\boldsymbol{X}$ is a discrete random variable. Then for any $\boldsymbol{x}$

$$\begin{aligned}
P_\theta(\boldsymbol{X}=\boldsymbol{x}) &= P_\theta(\boldsymbol{X}=\boldsymbol{x}, T(\boldsymbol{X})=T(\boldsymbol{x}))\ \text{, since } \{\boldsymbol{X}=\boldsymbol{x}\} \Rightarrow \{T(\boldsymbol{X})=T(\boldsymbol{x})\} \\
&= P(\boldsymbol{X}=\boldsymbol{x} \mid T(\boldsymbol{X})=T(\boldsymbol{x}))P_\theta(T(\boldsymbol{X})=T(\boldsymbol{x})) \\
&= P(\boldsymbol{Y}=\boldsymbol{x} \mid T(\boldsymbol{X})=T(\boldsymbol{x}))P_\theta(T(\boldsymbol{X})=T(\boldsymbol{x})) \\
&= P_\theta(\boldsymbol{Y}=\boldsymbol{x}, T(\boldsymbol{X})=T(\boldsymbol{x})) \\
&= P_\theta(\boldsymbol{Y}=\boldsymbol{x})
\end{aligned}$$

Consider the following examples.

Example 1: Let X be a single observation from $N(0,\sigma^2)$ distribution, where σ is unknown. Then, given $|X|=t$, the only two possible values of X are $+t$ or $-t$, and by symmetry, the conditional probability of each is $\frac{1}{2}$. The conditional distribution of X given $|X|=t$ is independent of σ and $T=|X|$ is sufficient for σ.

Note : Given $T=t$ we can generate a random variable Y which is equivalent to X by using the following pseudo random number generation technique:

We toss a fair coin and define a random variable Y such that

$$\begin{aligned}
Y &= +t \text{ if a head appears} \\
&= -t \text{ if a tail appears}
\end{aligned}$$

then the random variable Y is also a $N(0,\sigma^2)$ variable.

Example 2 : Let X_1, X_2 be i.i.d. $P(\lambda)$ random variables. Consider the statistic $T = X_1 + X_2$. Consider the conditional probability,

$$\begin{aligned}
P\left(X_1=x_1, X_2=x_2 | T=t\right) &= \frac{P(X_1=x_1, X_2=t-x_1)}{P(X_1+X_2=t)},\ \text{if } t=x_1+x_2, \\
&= 0,\ \text{otherwise.}
\end{aligned}$$

Thus, for $x_i = 0,1,2,\ldots, i=1,2, x_1+x_2=t$, we have

$$P\left(X_1=x_1, X_2=x_2 | X_1+X_2=t\right) = \binom{t}{x_1}\left(\frac{1}{2}\right)^t,$$

which is independent of λ. Hence $X_1 + X_2$ is sufficient for λ.

Note : Given $T = t$ we can generate a random vector (Y_1, Y_2) which is equivalent to (X_1, X_2) by using the following pseudo random number generation technique:

Toss a fair coin t times and let Y_1 and $Y_2 = t - Y_1$ be respectively the number of heads and number of tails obtained in t tosses. The the joint distribution of (Y_1, Y_2) is the same as that of (X_1, X_2).

Example 3: Let X_1, X_2 be independently and identically distributed as Poisson variables with parameter θ. Consider the statistic $T = X_1 + 2X_2$. The possible values of T are $0, 1, 2, \ldots$ We consider the conditional distribution of (X_1, X_2) given $T = t$ where $t = 0, 1, 2 \ldots$

$$\begin{aligned} P\left[X_1 = 0, X_2 = 0 | T = 0\right] &= \frac{P[X_0 = 1, X_2 = 0]}{P[X_1 + 2X_2 = 0]} \\ &= \frac{P[X_1 = 0, X_2 = 0]}{P[X_1 = 0, X_2 = 0]} \\ &= 1. \end{aligned}$$

$$\begin{aligned} P\left[X_1 = 1, X_2 = 0 | T = 1\right] &= \frac{P[X_1 = 1, X_2 = 0]}{P[X_1 + 2X_2 = 1]} \\ &= \frac{P[X_1 = 1, X_2 = 0]}{P[X_1 = 1, X_2 = 0]} \\ &= 1. \end{aligned}$$

$$\begin{aligned} P\left[X_1 = 0, X_2 = 1 | T = 2\right] &= \frac{P[X_1 = 0, X_2 = 1]}{P[X_1 + 2X_2 = 2]} \\ &= \frac{P[X_1 = 0, X_2 = 1]}{P[X_1 = 0, X_2 = 1] + P[X_1 = 2, X_2 = 0]} \\ &= \frac{\theta \exp\left(-2\theta\right)}{\theta \exp\left(-2\theta\right) + \frac{\theta^2}{2} \exp\left(-2\theta\right)}. \\ &= \frac{2}{2 + \theta}. \end{aligned}$$

and $P\left[X_1 = 2, X_2 = 0 | T = 2\right] = \frac{\theta}{2+\theta}$.

Since the conditional distribution of (X_1, X_2) given $T = 2$ is depends on θ, T is not

A Gateway to All Post Graduate Courses

sufficient for θ.

Example 4: Let X_1, X_2 be independently and identically distributed as Bernoulli variables with parameter θ. Consider the statistic $T = X_1 + 2X_2$.
The possible values of T are $0, 1$ and 2.

$$\begin{aligned} \{X_1 + 2X_2 = t\} &\iff \{X_1 = 0, X_2 = 0\}, \text{ if } t = 0 \\ &\iff \{X_1 = 1, X_2 = 0\}, \text{ if } t = 1 \\ &\iff \{X_1 = 0, X_2 = 1\}, \text{ if } t = 2 \end{aligned}$$

Hence we get

$$P\left[X_1 = 0, X_2 = 0 | T = 0\right] = 1$$
$$P\left[X_1 = 1, X_2 = 0 | T = 1\right] = 1$$
$$P\left[X_1 = 0, X_2 = 1 | T = 2\right] = 1$$

Since the conditional distribution of (X_1, X_2) given $T = t$ is independent of θ for every possible values of t, T is sufficient for θ.

Example 5: Let X_1, X_2 be independently and identically distributed with p.m.f.

$$\begin{aligned} f_N(x) &= \frac{1}{N}, \text{ if } x = 1, 2, \ldots, N \\ &= 0, \text{otherwise} \end{aligned}$$

where N is unknown and $N \in \{1, 2, \ldots\}$. Consider the statistic $T = Max(X_1, X_2)$. The possible values of T are $1, 2, \ldots N$. For any such mass point t we have

$$P_N(T = t) = P_N(T \leq t) - P_N(T \leq t-1) = \frac{1}{N^n}\left[t^n - (t-1)^n\right]$$

The conditional distribution of (X_1, X_2) given $T = t$ is

$$\begin{aligned} P\left[X_1 = x_1, X_2 = x_2 | T = t\right] &= \frac{1}{t^n - (t-1)^n} \text{ if } Max(x_1, x_2) = t \\ &= 0, \text{otherwise} \end{aligned}$$

Since the conditional distribution of (X_1, X_2) given $T = t$ is independent of N for every possible values of t, T is sufficient for N.

But if we consider the statistic $T_1 = Min(X_1, X_2)$, the possible values of T_1 are $1, 2, \ldots N$. For any such mass point t_1 we have

$$P_N(T_1 = t_1) = P_N(T_1 \geq t_1) - P_N(T_1 \geq t_1 + 1) = \frac{1}{N^n}\left[(N - t_1 + 1)^n - (N - t_1)^n\right].$$

The conditional distribution of (X_1, X_2) given $T_1 = t_1$ is

$$\begin{aligned} P\left[X_1 = x_1, X_2 = x_2 | T = t\right] &= \frac{1}{(N - t_1 + 1)^n - (N - t_1)^n} \text{ if } Min(x_1, x_2) = t \\ &= 0, \text{otherwise} \end{aligned}$$

Since the conditional distribution of (X_1, X_2) given $T_1 = t_1$ depends of N , T_1 is not sufficient for N.

Example 6 : Let X_1, X_2 be independently and identically distributed $N(\theta, 1)$ random variables. Consider two statistics $T = X_1 + X_2$ and $T_1 = X_1 + 2X_2$. The joint distribution of (T, T_1) is bivariate normal distribution with parameters $(2\theta, 3\theta, 2, 5, \frac{3}{\sqrt{10}})$. The conditional distribution of T given $T_1 = t_1$ is normal with mean $\frac{1}{5}(3t_1 - \theta)$ and variance $\frac{1}{5}$, whereas the conditional distribution of T_1 given $T = t$ is normal with mean $\frac{3}{2}t$ and variance $\frac{1}{2}$. Since the conditional distribution of T given $T_1 = t_1$ depends on θ, T_1 is not sufficient for θ. As the conditional distribution of T_1 given $T = t$ is independent of θ, T is sufficient for θ.

Minimal sufficiency

Definition A sufficient statistic T is said to be minimal sufficient statistic for θ if,

(i) If T is sufficient for θ.

(ii) T is a function of every other sufficient statistic.

A minimal sufficient provides the greatest possible reduction of the data. Consider the following example:

Example : Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $N(\theta, 1)$ population. The following statistics are sufficient for the parameter θ.

$$T_1 = (X_1, X_2, \ldots X_n)$$

$$T_2 = (X_1 + X_2, X_3, \ldots X_n)$$

$$T_3 = (X_1 + X_2 + X_3, X_4 \ldots X_n)$$

$$\ldots$$

$$T_n = (X_1 + X_2 + \ldots + X_n)$$

Note that T_i is a function of T_{i-1} for every $i = 2, 3, \ldots n$. T_n provides most thorough reduction of data and hence it is minimal sufficient for Θ.

e-pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

We have seen that for any statistic T, if the conditional distribution of $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$ given $T = t$ is independent of θ for every possible values of t then statistic T will be an sufficient statistic for θ. But in any problem there are infinitely many statistics and therefore it is a difficult task to find out which of these statistics is sufficient for the parameter θ. In the present module first we shall discuss a very simple method for determining a sufficient statistics. This method was to due to Fisher and Neyman and it is popularly known as Fisher-Neyman factorization theorem.

Factorization theorem Let $(X_1, X_2, \ldots X_n)$ be an n-dimensional random vector with joint p.m.f/ p.d.f. $f_\theta(x_1, x_2, \ldots x_n)$ where $\theta \in \Theta \subset \mathcal{R}$, is unknown. A statistic $T(X_1, X_2, \ldots X_n)$ is sufficient for θ if and only if $f_\theta(x_1, x_2, \ldots x_n)$ can be factorized as

$$f_\theta(x_1, x_2, \ldots, x_n) = g_\theta(t(x_1, x_2, \ldots x_n))h(x_1, x_2, \ldots, x_n), \tag{1}$$

where $g_\theta(t)$ depends on θ and on $(x_1, x_2, \ldots x_n)$ only through $t(x_1, x_2, \ldots x_n)$ and $h(x_1, x_2, \ldots, x_n)$ is independent of θ. Note: Factorization theorem also holds when θ is a vector of parameters and T is a multidimensional random variable. In that case, we say that T is jointly sufficient for θ.

Proof: For the sake of simplicity we assume that $(X_1, X_2, \ldots X_n)$ be an n-dimensional discrete random vector with joint p.m.f. $f_\theta(x_1, x_2, \ldots x_n)$. Let us write $\mathbf{X} = (X_1, X_2, \ldots X_n)$ and $\mathbf{x} = (x_1, x_2, \ldots x_n)$.

Only if part Suppose T be sufficient for θ. Since $\{\mathbf{X} = \mathbf{x}\} \iff \{T(\mathbf{X}) = t\}$ if $T(\mathbf{x}) = t$, and we can write,

$$\begin{aligned} f_\theta(\mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x}) &= P_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t), \text{ if } T(\mathbf{x}) = t \\ &= P_\theta(T = t)P(\mathbf{X} = \mathbf{x}|T = t), \end{aligned}$$

provided that $P(\mathbf{X} = \mathbf{x}|T = t)$ is well defined. Since T is sufficient for θ, the conditional probability $P(\mathbf{X} = \mathbf{x}|T = t)$ is independent of θ. Let us write

$$h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T = t).$$

Setting,
$$g_\theta(t) = P_\theta(T = t).$$
we get $f_\theta(\mathbf{x}) = g_\theta(t)h(\mathbf{x})$.

<u>*If part*</u> Suppose that (1) holds. Therefore for fixed t we have,

$$\begin{aligned} P_\theta(T = t) &= \sum_{\mathbf{x}:T(\mathbf{x})=t} P_\theta(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}:T(\mathbf{x})=t} g_\theta(T(\mathbf{x}))h(\mathbf{x}) \\ &= g_\theta(t) \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x}). \end{aligned}$$

Suppose that $P_\theta(T = t) > 0$. Then,

$$P_\theta(\mathbf{X} = \mathbf{x}|T = t) = \frac{P_\theta(\mathbf{X} = \mathbf{x}, T = t)}{P_\theta(T = t)},$$

which is 0 when $T(\mathbf{X}) \neq t$ and $\frac{P_\theta(\mathbf{X}=\mathbf{x})}{P_\theta(T=t)}$ when $T(\mathbf{X}) = t$ Thus if $T\mathbf{X} = t$,

$$\frac{P_\theta(\mathbf{X} = \mathbf{x}, T = t)}{P_\theta(T = t)} = \frac{g_\theta(t)h(\mathbf{x})}{g_\theta(t)\sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})}$$

which is independent of θ. Therefore T is sufficient for θ.

<u>Remark 1</u> If T is sufficient for θ and $T^\star = \phi(T)$ where $\phi(.)$ is an one to one function. Then $T^\star$ is also sufficient for θ.
Proof : Since $\phi(.)$ is one to one, we can write $T = \phi^{-1}(T^\star) = \psi(T^\star)$, say. By factorization theorem (only if part) we can write

$$\begin{aligned} f_\theta(\mathbf{x}) &= g_\theta(t)h(\mathbf{x}) \\ &= g_\theta(\psi(t^\star))h(\mathbf{x}) \\ &= g^\star(t^\star)h(\mathbf{x}), \text{ say, } \quad \text{where } g^\star(.) = g(\psi(.)). \end{aligned}$$

Hence factorization theorem (if part) , $T^\star$ is sufficient statistic for θ.
Note : Every function of a sufficient statistic may not be sufficient. For example, if $X \sim N(\theta, 1), \theta \in \mathcal{R}$ then X is sufficient for θ but X^2 is not. However, X is sufficient for θ^2.

Remark 2 Due to factorization theorem, a necessary and sufficient condition for a statistic T to be sufficient is that for any fixed θ_1 and θ_2, the ratio $f_{\theta_1}(x)/f_{\theta_2}(x)$ is a function of $T(x)$ only.

Remark 3 If T_1 and T_2 are two distinct sufficient statistics for θ, then T_1 is a function of T_2.
Proof: By factorization theorem,

$$f_\theta(\mathbf{x}) = g_\theta(t_1)h_1(\mathbf{x}) = g_\theta(t_1)h_2(\mathbf{x})$$

which implies that T_1 is a function of T_2.

Remark 4 If (T_1, T_2) is jointly sufficient for (θ_1, θ_2), then it does not follow that T_i is sufficient for $\theta_i, i = 1, 2$.

Remark 5: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a continuous distribution with p.d.f. $f_\theta(x)$. Then the set of order statistics $\mathbf{T} = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is jointly sufficient for θ. Proof: The joint p.d.f. of $\mathbf{T}$ is given by

$$g_\theta(\mathbf{t}) = n! \; \Pi_{i=1}^n f_\theta(x_{(i)})$$

Since $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is nothing but a permutation of $(x_1, x_2, \ldots, x_n)$, we have

$$g_\theta(\mathbf{t}) = n! \; \Pi_{i=1}^n f_\theta(x_i)$$

$$\Rightarrow f_\theta(\mathbf{x}) = g_\theta(\mathbf{t})\frac{1}{n!}$$

Setting $h(\mathbf{x}) = \frac{1}{n!}$, the remark follows from the factorization theorem.

Some examples:

Example 1: Consider a set of n Bernoullian trials with probability of success p. With the i^{th} trial we may associate a variable X_i having the probability-mass function

$$f(x_i|p) = p^{x_i}(1-p)^{1-x_i}, \text{ for } x_i = 0, 1$$

The joint probability-mass function of $x_1, x_2, \ldots, x_n$ is,

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|p) &= p^{T(x_1,x_2,\ldots,x_n)}(1-p)^{(n-T(x_1,x_2,\ldots,x_n))} \\ &= g(T(x_1, x_2, \ldots, x_n)|p)h(x_1, x_2, \ldots, x_n), \end{aligned}$$

where $T(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$, $g(T(x_1, x_2, \ldots, x_n)|p) = p^{T(x_1,x_2,\ldots,x_n)}(1 - p)^{(n-T(x_1,x_2,\ldots,x_n))}$ and $h(x_1, x_2, \ldots, x_n) = 1$. Hence $T(x_1, x_2, \ldots, x_n)$, the total number of successes out of n trials, is a sufficient statistics for p. So is $T(x_1, x_2, \ldots, x_n)/n$, the average number of success.

Example 2: Let $x_1, x_2, \ldots, x_n$ be n independent random observations from a normal distribution with mean μ is unknown and variance $\sigma^2 < \infty$ is known. Then the joint density function of $x_1, x_2, \ldots, x_n$ is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \\ &= g(T(x_1, x_2, \ldots, x_n)|\mu)h(x_1, x_2, \ldots, x_n), \end{aligned}$$

where $T(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$ as well as $\bar{x}$ is a sufficient statistic for μ.

Example 3: Let $x_1, x_2, \ldots, x_n$ be an independent random sample from a Poisson distribution with parameter $\lambda > 0$ unknown and the corresponding probability mass function is

$$f(x_i|\lambda) = \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \text{ for } x_i = 0, 1, 2, \ldots, \infty.$$

The joint density of $x_1, x_2, \ldots, x_n$ is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu) &= \frac{\exp(-n\lambda)\lambda^{T(x_1,x_2,\ldots,x_n)}}{\prod_{i=1}^{n} x_i!} \\ &= g(T(x_1, x_2, \ldots, x_n)|\lambda)h(x_1, x_2, \ldots, x_n), \end{aligned}$$

where $T(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$ as well as $\bar{x}$ is a sufficient statistic for λ.

Example 4: Let $x_1, x_2, \ldots, x_n$ be independently and identically distributed as Uniform $[0, \theta]$. The corresponding density for each x_i is,

$$f(x_i|\theta) = \theta^{-1} I\{0 < x_i < \theta\},$$

$I\{A\}$ is used to denote the indicator function of set A. The the joint density of $x_1, x_2, \ldots, x_n$ is

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^{-n} I\{0 < x_i < \theta \; \forall i = 1, 2, \ldots, n\}$$

$$
\begin{aligned}
&= \theta^{-n} I\{0 < T(x_1, x_2, \ldots, x_n) < \theta\} \\
&= g(T(x_1, x_2, \ldots, x_n)|\theta) h(x_1, x_2, \ldots, x_n),
\end{aligned}
$$

where $T(x_1, x_2, \ldots, x_n) = \max_{i=1}^{n} x_i$ is the sufficient statistic for θ.

Example 5: Let $X_1, X_2, \ldots, X_n$ be n independent random observations from a normal distribution with mean μ and variance $\sigma^2 < \infty$ both unknown. Then the joint density function of $X_1, X_2, \ldots, X_n$ is

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n|\mu, \sigma) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \\
&= \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\sigma} \exp\left(-\frac{n}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu}{2\sigma^2}\right)
\end{aligned} \tag{2}
$$

Here $\theta = (\mu, \sigma^2)$ is a vector valued parameter. Therefore, $T(X_1, X_2, \ldots, X_n) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is jointly sufficient statistic for μ and σ^2.

Example 6: Let $X_1, X_2, \ldots, X_n$ be n independent random observations from $U(\theta, \theta + 1)$ where θ unknown. Then the joint density function of $X_1, X_2, \ldots, X_n$ is

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n|\theta) &= I\left(\theta \le x_1, x_2, \ldots, x_n \le \theta + 1\right) \\
&= I\Big(\theta \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \le \theta + 1\Big) \\
&= I\Big(\theta \le x_{(1)} \le x_{(n)} \le \theta + 1\Big).
\end{aligned}
$$

Therefore, $T(X_1, X_2, \ldots, X_n) = \Big(X_{(1)}, X_{(n)}\Big)$ is jointly sufficient statistic for θ.

Subject-Statistics

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Exponential families

The exponential family of distributions happens to be very rich when it comes to statistical modeling of data sets in practice. The distributions belonging to this family enjoy many interesting properties which often attract investigators toward specific members of this family in order to pursue statistical studies. In the present module we discuss briefly both the one-parameter and multi-parameter exponential family of distributions. Some of those properties and underlying data reduction principles, such as sufficiency or completeness, are discussed in subsequent modules.

One-parameter exponential families

Definition Let $\mathcal{F}_\theta = \{f_\theta(x), x \in \mathcal{X}, \theta \in \Theta\}$ be a family of probability density or mass functions. The parameter space Θ is an open subinterval of real line and the support $\mathcal{X} = \{x : f_\theta(x) > 0\}$ is independent of θ. We say that it is an one-parameter exponential family if and only if we can express $f_\theta(x)$ as,

$$f_\theta(x) = \exp\left[q(\theta)T(x) + u(\theta) + v(x)\right], \tag{1}$$

for some real valued functions $q(\theta)$ and $u(\theta)$ of θ and $T(x)$ and $v(x)$ of x.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with density or mass function f_θ. Then the joint density of $X_1, X_2, \ldots, X_n$ is given by,

$$g_\theta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \exp\left\{q(\theta)\sum_{i=1}^{n} T(x_i) + u(\theta) + \sum_{i=1}^{n} v(x_i)\right\}.$$

Hence, g_θ is also of the form (1).

Some examples:

Example 1: Let $X \sim N(0, \sigma^2)$ where, standard deviation σ^2 is unknown. Then,

$$f_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left(\frac{x^2}{2\sigma^2}\right) = \exp\left(-\frac{\log(2\pi\sigma^2)}{2} - \frac{x^2}{2\sigma^2}\right),$$

is a one-parameter exponential family with

$$q(\sigma^2) = -\frac{1}{2\sigma^2}, \quad T(x) = x^2, \quad u(\sigma^2) = -\frac{\log(2\pi\sigma^2)}{2}, \quad v(x) = 0.$$

Example 2: Let $X \sim N(\mu, 1)$ where, mean μ is unknown. Then,

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp - \left(\frac{(x-\mu)^2}{2}\right) = \exp\left(-\frac{\log(2\pi)}{2} - \frac{x^2}{2} - \frac{\mu^2}{2} + x\mu\right),$$

is a one-parameter exponential family with

$$q(\mu) = \mu, \quad T(x) = x, \quad u(\mu) = -\frac{\mu}{2}, \quad v(x) = -\left[\frac{x^2}{2} + \frac{\log(2\pi)}{2}\right].$$

Example 3: Let $X \sim P(\lambda)$ where, $\lambda > 0$ is unknown. Then the probability mass function is,

$$P_\lambda(X = x) = \exp(-\lambda)\frac{\lambda^x}{x!} = \exp(-\lambda + x\log(\lambda) - \log(x!)),$$

is a member of one-parameter exponential family with

$$q(\lambda) = \log(\lambda), \quad T(x) = x, \quad u(\lambda) = -\lambda, \quad v(x) = -\log(x!).$$

Example 4: Let $X \sim Uniform(0, \theta)$. Then the probability density function is given by,

$$f_\theta(x) = \frac{1}{\theta} \text{ if, } 0 < x < \theta$$

. Since the support of $f_\theta(x)$ depends on θ, it does not belong to the one-parameter exponential family.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Example 5: Let $X \sim N(\theta, \theta^2), \theta > 0$. Then the probability density function is given by,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\theta^2}} \exp - \left(\frac{(x-\theta)^2}{2\theta^2} \right) = \exp \left(-\frac{\log(2\pi)}{2} - \frac{x^2}{2\theta^2} - \frac{1}{2\theta} + \frac{2x}{\theta} \right),$$

Since $f_\theta(x)$ can not be expressed in the form (1), it does not belong to the one-parameter exponential family.

Multi-parameter exponential families

Now we would like to introduce the multi-parameter exponential families with k-parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_k)$.

Definition Let $\mathcal{F}_{\boldsymbol{\theta}} = \{f_{\boldsymbol{\theta}}(x), x \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ be a family of probability density or mass functions. The parameter space Θ is an open set of $\mathcal{R}^k$ and the support $\mathcal{X} = \{x : f_{\boldsymbol{\theta}}(x) > 0\}$ is independent of $\boldsymbol{\theta}$. We say that it is a k-parameter exponential family if and only if we can express $f_{\boldsymbol{\theta}}(x)$ as,

$$f_{\boldsymbol{\theta}}(x) = \exp \left[\sum_{j=1}^{k} q_i(\boldsymbol{\theta}) T_i(x) + u(\theta) + v(x) \right], \tag{2}$$

for some real valued functions $q_i(\theta)$ and $u_i(\theta), i = 1, 2, \dots k$ of θ and $T_i(x), i = 1, 2, \dots k$ and $v(x)$ of x.

Some examples:

Example 1: Let $X \sim N(\mu, \sigma^2)$ when both μ and σ^2 are unknown. We have,

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} \right) = \exp \left\{ -\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x - \frac{1}{2} \left[\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2) \right] \right\}$$

is a two-parameter exponential family with

$$q_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2}, \; q_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, \quad T_1(x) = x^2, \; T_2(x) = x,$$

$$u(\mu, \sigma^2) = -\frac{1}{2} \left[\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right], \quad v(x) = 0.$$

Example 2: Let $X \sim Beta(\theta_1, \theta_2)$ with both θ_1 and θ_2 unknown. We have,

$$f_{\theta_1,\theta_2}(x) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1+\theta_2)} x^{\theta_1-1}(1-x)^{\theta_2-1} \text{ if, } 0 < x < 1, \theta_1, \theta_2 > 0.$$

Then,

$$f_{\theta_1,\theta_2}(x) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1+\theta_2)} \exp[(\theta_1-1)\log(x) + (\theta_2-1)\log(1-x)] I\left\{0 < x < 1\right\},$$

where, where $I(A)$ is the indicator function for the set A. Therefore, $Beta(\theta_1, \theta_2)$ belongs to a two-parameter exponential family with,

$$q_1(\theta_1, \theta_2) = (\theta_1 - 1),\ Q_2(\theta_1, \theta_2) = (\theta_2 - 1),\ T_1(x) = \log(x),\ T_2(x) = \log(1-x),$$

$$u(\theta_1, \theta_2) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1+\theta_2)},\quad v(x) = I\left\{0 < x < 1\right\}.$$

Example 3 : Let $X \sim Rect(\theta_1, \theta_2)$ with both θ_1 and θ_2 unknown. We have,

$$f_{\theta_1,\theta_2}(x) = \frac{1}{\theta_2 - \theta_1}, \theta_1 < x < \theta_2.$$

Since the support of the distribution depends on (θ_1, θ_2), the p.d.f. $f_{\theta_1,\theta_2}(x)$ does not belong to two=parameter exponential family.

Example 4 : Let $X \sim Exp(\mu, \sigma)$ with both μ and σ unknown. We have,

$$f_{\mu,\sigma} = \frac{1}{\sigma} e^{-(\frac{x-\mu}{\sigma})}, x > \mu, -\infty < \mu < \infty, \sigma > 0$$

Since the support of the distribution depends on μ, the p.d.f. $f_{\theta_1,\theta_2}(x)$ does not belong to two-parameter exponential family.

In all the above examples we consider a single random variable X. But the concept of exponential family can also be extended to the joint distribution of correlated random vectors of dimension ≥ 2. For example,

(i) If $(X_1, X_2) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then the joint distribution of (X_1, X_2) belongs to the five-parameter exponential family.

(ii) If $(X_1, X_2, \ldots, X_p) \sim N_p(\mu, \Sigma)$, then the joint distribution of $(X_1, X_2, \ldots, X_p)$ belongs to the k-parameter exponential family where $k = p + \frac{p(p+1)}{2} = \frac{p(p+3)}{2}$.

Now we consider the following very important result.

Result If X is belongs to exponential family, and has the form,

$$f_\theta(x) = \exp\left[q(\theta)t(x) + u(\theta) + v(x)\right]$$

then, $T = t(x)$ also belongs to an exponential family.

Proof. Let us define the family of distribution (discrete case) for T by,

$$\begin{aligned}
f_\theta^T(t) &= \sum_{x:T(x)=t} f_\theta(x) \\
&= \sum_{x:T(x)=t} \exp\left[Q(\theta)T(x) + D(\theta) + S(x)\right] \\
&= \exp[D(\theta)] \sum_{x:T(x)=t} \exp\left[Q(\theta)T(x) + S(x)\right] \\
&= \exp\left[Q(\theta)t + D(\theta)\right] \sum_{x:T(x)=t} \exp\{S(x)\} \\
&= \exp\left[Q(\theta)t + D(\theta) + H(t)\right],
\end{aligned}$$

where $H(x) = \sum_{x:T(x)=t} \exp\{S(x)\}$. Hence by definition of exponential family T also belongs to one.

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Completeness

Definition Consider a family of probability density or mass functions $\mathcal{F}_\theta = \{f_\theta(x) : \boldsymbol{\theta} \in \Theta\}$. Consider any real valued function $\psi(x)$ having finite expectation. We say that the family is complete if,

$$E_\theta(\psi(X)) = 0 \quad \forall\ \theta \in \Theta$$

implies that,

$$P_\theta\left(\psi(X) = 0\right) = 1 \quad \forall\ \theta \in \Theta.$$

The family $\mathcal{F}_\theta$ is boundedly complete if for any bounded function $\psi(X)$ of X

$$E_\theta(\psi(X)) = 0 \quad \forall\ \theta \in \Theta$$

implies that,

$$P_\theta\left(\psi(X) = 0\right) = 1 \quad \forall\ \theta \in \Theta.$$

Let T be a statistic and $\mathcal{F}_\theta^T = \{f_\theta^T(x) : \boldsymbol{\theta} \in \Theta\}$ be the family of probability distributions of T.

Definition The statistic $T(X)$ is said to be a complete (boundedly complete) if, the family of distribution of T is complete (boundedly complete)i.e. for any real valued function $\psi(t)$ having finite expectation

$$E_\theta(\psi(T)) = 0 \quad \forall\ \theta \in \Theta$$

implies that,

$$P_\theta\left(\psi(T) = 0\right) = 1 \quad \forall\ \theta \in \Theta.$$

Implication A statistic T is complete means that there exists no non-trivial function of T which is an unbiased estimator of 0.

Result 1 : If T is complete and $T^\star$ is a function of T, then $T^\star$ is complete.

Proof : Let $T^\star = \phi(T)$ and $\psi(T^\star)$ be any real valued function of $t^\star$

$$E_\theta(\psi(T^\star)) = 0 \quad \forall\ \theta \in \Theta$$

$$\Rightarrow E_\theta(\psi(\phi(T))) = 0 \quad \forall\ \theta \in \Theta$$

$$\Rightarrow \psi(\phi(T)) = 0 \quad \forall\, \theta \in \Theta,\ \text{since } T \text{ is complete}$$

$$\Rightarrow \psi(T^{\star})) = 0 \quad \forall\, \theta \in \Theta,\ \text{since } T \text{ is complete}$$

Hence $T^{\star}$ is complete.
Result 2 : If T is complete then T is boundedly complete.
Proof : If $E_{\theta}(\psi(T)) = 0 \quad \forall\, \theta \in \Theta$ implies that, $P_{\theta}\left(\psi(T) = 0\right) = 1 \quad \forall\, \theta \in \Theta$ for any function $\psi(T)$. Then it must also be true for the subset of bounded functions.
Note : The converse of this result is not true.

Counter example : Let X be a random variable with p.m.f

$$\begin{aligned} f_{\theta}(x) &= \theta,\ \text{if } x = -1 \\ &= (1-\theta)^2\theta^x,\ \text{if } x = 0, 1, 2, \ldots, 0 < \theta < 1. \end{aligned}$$

Let $\psi(X)$ be any function of X.

$$E_{\theta}(\psi(X)) = 0 \ \text{ for all } \ \theta$$

$$\Rightarrow \theta\psi(-1) + \sum_{x=0}^{\infty} \psi(x)(1-\theta)^2\theta^x = 0 \ \text{ for all } \ \theta$$

$$\Rightarrow \sum_{x=0}^{\infty} \psi(x)\theta^x = \frac{-\theta\psi(-1)}{(1-\theta)^2} = -\sum_{x=0}^{\infty} \psi(-1)x\theta^x \ \text{ for all } \ \theta$$

Comparing the coefficients of $\theta^x, x = 0, 1, 2, \ldots$ from both sides of the above identity we get

$$\psi(x) = -xh(-1),\ \text{for all } x$$

If $\psi(-1) = c \neq 0$, then $\psi(x) = -cx$, for all x, i.e. $\psi(x)$ is non-zero with non-zero probability.
Hence the family of distributions $\mathcal{F}_{\theta} = \{f_{\theta}(x) : \boldsymbol{\theta} \in \Theta\}$ is not complete.
Again $\psi(x)$ is bounded if and only if $\psi(-1) = 0$ in which case $\psi(x) = 0$, for all $x = 0, 1, 2 \ldots$. Hence the family of distributions $\mathcal{F}_{\theta} = \{f_{\theta}(x) : \boldsymbol{\theta} \in \Theta\}$ is complete.

Some examples:

Example 1: Let $X \sim Binomial(n, \theta)$ random variables. Consider any real valued $\psi(x)$ such that

$$E_\theta(\psi(X)) = 0 \ \ \forall 0 < \theta < 1,$$

$$\Rightarrow \sum_{x=0}^{n} \psi(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = 0 \quad \forall 0 < \theta < 1,$$

If we set $\lambda = \frac{\theta}{1-\theta}$ then

$$\sum_{x=0}^{n} \psi(t) \binom{n}{x} \lambda^x = 0 \quad \forall \lambda \in (0, \infty),$$

Which is a polynomial in λ of degree n. The only way a polynomial of degree n can be zero for more than n distinct values of the variable is for all coefficients to be zero. That is,

$$\psi(x) = 0 \text{ for } x = 0, 1, \ldots, n$$

Thus, we have shown that, if $E_\theta(\psi(X)) = 0 \ \ \forall \ \theta \in (0,1)$, then $\psi(X) = 0$ with probability one. Hence the binomial complete is complete.
Remark 1: If the parameter space Θ is not an open interval of real line, for example, if $\Theta = \{1/3, 2/3\}$ then the family of distributions of T is not complete.

Remark 2 : Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(θ) random variables. Consider the statistic $T(X_1, X_2, \ldots, X_n) = \sum\limits_{i=1}^{n} X_i$ ~Binomial(n, θ). Then T has a complete family of distributions and hence T is complete.
The joint p;m.f. of $(X_1, X_2, \ldots, X_n)$ is given by

$$f_\theta(x_1, x_2, \ldots, x_n) = \theta^{\sum\limits_{i=1}^{n} X_i} (1-\theta)^{\sum\limits_{i=1}^{n}}, x_i = 0, 1 \text{ for all } \ i = 1, 2 \ldots n$$

The family of joint distributions of $(X_1, X_2, \ldots, X_n)$ is not complete since if we take $\psi((X_1, X_2, \ldots, X_n)) = X_1 - X_2$ then

$$E_\theta(\psi(X_1, X_2, \ldots, X_n)) = E_\theta(X_1 - X_2) = 0 \quad \forall \ \theta \in (0,1)$$

but $P_\theta(X_1 - X_2 \neq 0) > 0$.

Example 2: Let $X_1, X_2, \ldots, X_n$ be i.i.d with common p.m.f.

$$f_N(x) = \frac{1}{N}, x = 1, 2, \ldots N, N \text{ is a positive integer}$$

Let $\psi(X)$ be any real valued function of X such that

$$E_N(\psi(X)) = 0 \ \text{ for all } \quad N \geq 1$$

$$\frac{1}{N}\sum_{x=1}^{N}\psi(X) = 0 \ \text{ for all } \quad N \geq 1,$$

and this happens if and only if $\psi(x) = 0$ for all $x = 1, 2, \ldots N$. Hence the family $\mathcal{F}_N = \{f_N(x) : N \geq 1\}$.
Now suppose that we exclude the value $N = n_0$ for some fixed $n_0 \geq 1$ from the family $\mathcal{F}_N = \{f_N(x) : N \geq 1\}$. Let us write $\mathcal{F}_N^\star = \{f_N(x) : N \geq 1, N \neq n_0\}$. The family $\mathcal{F}_N^\star$ is not complete. To show this we consider the function

$$\begin{aligned}
\phi(X) &= 0, \ \text{ if } \ x = 1, 2, \ldots, n_0 - 1, n_0 + 2, n_0 + 3, \ldots \\
&= c \ \text{ if } \ x = n_0 \\
&= -c \ \text{ if } \ x = n_0 + 1,
\end{aligned}$$

where c is a constant,$c \neq 0$.
Then $E_N(\phi(X)) = 0$ for all $N \geq 1, N \neq n_0$ but function $P_N(\phi(X) = 0) \neq 1$. Hence the family $\mathcal{F}_N^\star$ is not complete.
Note: This example shows that the exclusion of even one member from the family $\mathcal{F}_N$ destroys completeness.
Example 3 : Let $X \sim N(\theta, 1)$. Consider any real valued $\psi(x)$ such that

$$E_\theta(\psi(X)) = 0 \ \ \forall \theta \in \mathcal{R}$$

$$\Rightarrow \int_{-\infty}^{\infty} \psi(x)\frac{1}{\sqrt{2\pi}}e^{-1/2(x-\theta)^2} = 0 \ \forall \ \theta$$

$$\Rightarrow \int_{-\infty}^{\infty} \psi(x)e^{\frac{-x^2}{2}}e^{\theta x}dx = 0 \ \forall \ \theta$$

Since integral is a bilinear laplace transform of $\psi(x)e^{\frac{-x^2}{2}}$, we get

$$\psi(x)e^{\frac{-x^2}{2}} = 0 \ \forall x$$

$$\Rightarrow \psi(x) = 0 \ \forall x$$

Hence $N(\theta, 1)$ family is complete.
Example 4 : Let $X \sim N(0, \theta^2)$. If we take $\psi(x) = x$ then

$$E_\theta(\psi(X)) = 0 \quad \forall \theta > 0$$

but the function $\psi(X) \neq 0$ with probability one.Hence $N(0, \theta^2)$ family is complete.
Example 5: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random sample from Normal distribution with mean 0, and the variance σ unknown. Here the statistic $T = \sum_{i=1}^n X_i^2 \sim \sigma\chi_n^2$ and the density function is given by,

$$g_\sigma(t) = \frac{1}{(2\sigma)^{n/2}\Gamma(n/2)} \exp -(t/2\sigma) t^{(n-2)/2} \quad 0 < t < \infty.$$

Now for any measurable function $\psi(t)$,

$$E_\sigma(\psi(T)) = \frac{1}{(2\sigma)^{n/2}\Gamma(n/2)} \int_0^\infty \psi(t) \exp -(t/2\sigma) t^{(n-2)/2} dt = 0 \quad \forall \sigma \in (0, \infty),$$

and on using Laplace transforms, it follows that, $\psi(t)t^{(n-2)/2} = 0$ and *i.e.* $\psi(t) = 0$ almost everywhere. Hence T is a complete statistic for σ.
Example 6: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random observations from a uniform distribution $U(0, \theta)$. The density function for each X_i is

$$\begin{aligned} f_\theta(x_i) &= \frac{1}{\theta} \text{ if } 0 < x_i < \theta \\ &= 0 \end{aligned}$$

Here $T = X_{(n)}$, the largest order statistics is a sufficient statistic. Moreover we will show it's completeness. Let us denote the probability density function of T by $g_\theta(t)$ and is given by,

$$\begin{aligned} g_\theta(t) &= nt^{n-1}/\theta^n \text{ if } 0 < t < \theta \\ &= 0 \text{ o.w.} \end{aligned}$$

Now, for a measurable function $\psi(T)$,

$$E_\theta(\psi(T)) = \frac{n}{\theta^n} \int_0^\theta \psi(t) t^{n-1} dt = 0 \quad \forall \theta \in (0, \infty),$$

or,

$$\int_0^\theta \psi(t) t^{n-1} dt = 0 \quad \forall \theta \in (0, \infty)$$

Differentiating both sides with respect to θ gives,$\psi(\theta) = 0,\ \forall\theta \in (0,\infty)$ and hence $\psi(t) = 0,\ \forall t \in (0,\infty)$. Since $(0,\theta) \subset (0,\infty)$, we get $\psi(t) = 0,\ \forall t \in (0,\theta)$. Hence $T = X_{(n)}$ is a complete statistic.

Example 7: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random sample from Normal distribution with mean μ, and variance σ both unknown. We know that $T = (T_1, T_2)$, where

$$T_1 = \sqrt{n}\bar{X},\ \text{and}\ T_2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

is a sufficient statistics for $\theta = (\mu, \sigma^2)$. Now the joint density function of T_1, T_2 is

$$f_\theta(t_1, t_2) = C\exp\left[-\frac{(t_1 - \sqrt{n}\mu)^2}{2\sigma^2}\right] t_2^{(n-3)/2} \exp\left[-\frac{t_2}{2\theta_2}\right].$$

Suppose now that the function $\psi(T)$ is such that

$$E_\theta(\psi(T)) = \int_{-\infty}^{\infty}\int_0^{\infty} \psi(t_1, t_2) f_\theta(t_1, t_2) dt_1 dt_2 = 0,\ \text{for all}\ \theta \in \Theta,$$

i.e.

$$\int_{-\infty}^{\infty}\int_0^{\infty} \psi(t_1, t_2)\exp\left[-\frac{1}{2\sigma^2}(t_1^2 + t_2) + \frac{\sqrt{n}\mu t_1}{\sigma^2}\right] t_2^{(n-3)/2} dt_1 dt_2 = 0,\ \text{for all}\ \theta \in \Theta.$$

Changing the variable t_2 to $u = t_1^2 + t_2$, then this reduces to

$$\int_{-\infty}^{\infty}\int_{t_1^2}^{\infty} \phi_(t_1, u)\exp\left[-\frac{u}{2\sigma^2} + \frac{\sqrt{n}\mu t_1}{\sigma^2}\right] \xi(t_1, u) du dt_1 = 0,\ \text{for all}\ \theta \in \Theta.$$

where $\phi[t_1, u(t_1, t_2)] = \psi(t_1.t_2)$ and $\xi(t_1, u) = (u - t_1^2)^{(n-3)/2} \neq 0$ for $u > t_1^2$. If we now extend the range of $\xi(t_1, u)$ by taking $\xi(t_1, u) = 0$ for $0 < u \leq t_1^2$, then the above identity becomes

$$\int_{-\infty}^{\infty}\int_0^{\infty} \phi_(t_1, u)\exp\left[-\frac{u}{2\sigma^2} + \frac{\sqrt{n}\mu t_1}{\sigma^2}\right] \xi(t_1, u) du dt_1 = 0,\ \text{for all}\ \theta \in \Theta.$$

Let us write $\frac{1}{2\sigma^2} = \theta_1'$ and $-\frac{\sqrt{n}\mu}{\sigma^2} = \theta_2'$, so that the above is seen to be a Laplace transform. From the uniqueness property of the Laplace transform, it follows that

$$\phi(t_1, u) = 0 \text{ for almost all } (t_1, u), u > 0,$$

i.e.,

$$\psi(t_1, t_2) = 0 \text{ for almost all } (t_1, t_2), t_2 > 0,$$

Hence $T = (T_1, T_2)$ has a complete family of distribution.

On other hand when we will be needing to show that a statistic T is not complete statistic for some parameter θ, all we will need is to search for some measurable function $\psi^*(t)$ such that

$$E_\theta(\psi^*(T)) = 0 \quad \forall \theta \in \Theta,$$

but $\psi^*(t) \neq 0$ almost everywhere. Following examples will clear this approach.

Example 8: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random sample from $N(\theta, \theta^2)$. Then $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is jointly sufficient for θ. However T is not complete since,

$$E_\theta\left[2\left(\sum_{i=1}^n X_i\right)^2 - (n+1)\sum_{i=1}^n X_i^2\right] = 0 \quad \text{for all } \theta$$

but, $\psi^*(T) = 2\left(\sum_{i=1}^n X_i\right)^2 - (n+1)\sum_{i=1}^n X_i^2$ is not identically zero.

Example 9: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random sample from $U(\theta, \theta+1)$. Then $T = (X_{(1)}, X_{(n)})$ is jointly sufficient for θ. However T is not complete since,

$$E_\theta\left[X_{(n)} - X_{(1)} - \frac{n-1}{n+1}\right] = 0 \quad \text{for all } \theta$$

but, $\psi^*(T) = X_{(n)} - X_{(1)} - \frac{n-1}{n+1}$ is not identically zero.

Minimal sufficiency

Definition A sufficient statistic T is said to be minimal if all sufficient statistic it provides the greatest possible reduction of the data, that is, if for any other sufficient statistic U there exist a function H such that $T = H(U)$ almost everywhere.

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Complete Sufficient Statistic

Definition A statistic T is called complete sufficient for an unknown parameter θ if and only if T is sufficient for θ and T is complete.

Now we shall consider the problem of determining complete sufficient statistic for exponential family of distributions. In this connection let us consider the following results.First we consider the case of an one parameter exponential family.

Result 1 : If be the p.m.f. or p.d.f. of a random variable X is of the form

$$f_\theta(x) = \exp\left[q(\theta)t(x) + u(\theta) + v(x)\right],$$

then the statistic $T(X)$ is complete sufficient for θ.

Proof : We can factorize $f_\theta(x)$ as

$$f_\theta(x) = g_\theta(t)h(x),$$

where $g_\theta(t) = u(\theta)\exp\left[q(\theta)t(x)\right]$ and $h(x) = \exp\left[v(x)\right]$.

Hence by factorization theorem T is sufficient for θ.

To prove the completeness without loss of generality we assume that $f_\theta(x)$ is continuous in $\mathcal{R}$. Now the p.d.f. of T can be written as

$$f^\star(t) = exp[u(\theta)]\exp\left[q(\theta)t(x)\right]h^\star(t)$$

Let $\psi(T)$ be any real valued function of T such that

$$E_\theta(\psi(T)) = 0, \text{for all } \theta$$

$$\Rightarrow exp[u(\theta)]\int_{-\infty}^{\infty}\psi(t)\exp\left[q(\theta)t(x)\right]h^\star(t) = 0, \text{ for all } \theta$$

Since the integral is of the Laplace transform type, it follows that

$$\psi(t)h^\star(t) = 0, \text{ for all} t \in \mathcal{R}$$

Since $h^\star(t) \neq 0$ for all t we get

$$\psi(T) = 0 \text{ with probability one}$$

Hence T is complete.

Note : $X_1, X_2, \ldots, X_n$ be n be a random sample from the above population $f_\theta(x)$, then $\sum\limits_{i=1}^{n} T_i$ is complete sufficient for θ.

Let us consider some examples.

Example 1 : If $X_1, X_2, \ldots, X_n$ be n be a random sample from $P(\theta)$ population $\sum\limits_{i=1}^{n} X_i$ is complete sufficient for θ.

Example 2 : If $X_1, X_2, \ldots, X_n$ be n be a random sample from $NB(r, \theta)$ population, where r is known and $\theta \in (0, 1)$ is unknown, $\sum\limits_{i=1}^{n} X_i$ is complete sufficient for θ.

Example 3 : If $X_1, X_2, \ldots, X_n$ be n be a random sample from $Beta(\theta, 1)$ population $\prod\limits_{i=1}^{n} X_i$ is complete sufficient for θ.

Example 4 : f $X_1, X_2, \ldots, X_n$ be n be a random sample from $N(\theta, 1)$ population $\sum\limits_{i=1}^{n} X_i$ is complete sufficient for θ.

Example 5 : f $X_1, X_2, \ldots, X_n$ be n be a random sample from $N(0, \theta^2)$ population $\sum\limits_{i=1}^{n} X_i^2$ is complete sufficient for θ.

In order to determine complete sufficient statistic for a k-parameter exponential family we consider the following result.

Result 2 : If be the p.m.f. or p.d.f. of a random variable X is of the form

$$f_{\boldsymbol{\theta}}(x) = \exp\left[\sum_{j=1}^{k} q_i(\theta) T_i(x) + u(\theta) + v(x)\right],$$

then the statistic $(T_1, T_2, \ldots T_k)$ is complete sufficient for $(\theta_1, \theta_2, \ldots \theta_k)$.

Now we discuss some examples of determining complete sufficient statistic where the distribution does not belong to the exponential family.

Example 1 : Let $X_1, X_2, \ldots, X_n$ be n be a random sample from the population with p.m.f.

$$f_N(x) = \frac{1}{N}, x = 1, 2, \ldots N, N(\geq 1) \text{ is a positive integer.}$$

The joint p.m.f. of $X_1, X_2, \ldots, X_n$ is given by

$$f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n) = \frac{1}{N^n}, \text{ if } 1 \leq x_{(n)} \leq N$$

Let us define a function

$$\begin{aligned}\phi(a,b) &= 1 \text{ if } a \le b \\ &= 0 \text{ if } a > b\end{aligned}$$

then $f_N(x_1, x_2, \dots x_n)$ can be written as

$$f_{\boldsymbol{\theta}}(x_1, x_2, \dots x_n) = \frac{1}{N^n}\phi(x_{(n)}, N)$$

If we write $t = x_{(n)}$, $g_N(t) = \frac{1}{N^n}\phi(t, N)$ and $h(x_1, x_2, \dots x_n) = 1$, then

$$f_{\boldsymbol{\theta}}(x_1, x_2, \dots x_n) = g_N(t) = \frac{1}{N^n}\phi(t, N)h(x_1, x_2, \dots x_n)$$

By factorization theorem, $T = X_{(n)}$ is sufficient for N.
The p.m.f. of T is given by

$$f_N^\star(t) = \frac{1}{N^n}\left[t^n - (t-1)^n\right], t = 1, 2, \dots N$$

Let $\psi(T)$ be any real valued function of T such that

$$E_N(\psi(T)) = 0, \text{ for all} N \ge 1$$

$$\Rightarrow \sum_{t=1}^{N}\psi(t)\frac{1}{N^n}\left[t^n - (t-1)^n\right]0, \text{ for all} N \ge 1 \tag{1}$$

Since (1) holds for all values of $N \ge 1$, setting $N = 1$ in (1) we get

$$\psi(1) = 0$$

Setting $N = 2$ in (1) we get

$$\psi(1) + \psi(2)(2^n - 1) = 0 be$$

$$\Rightarrow \psi(2) = 0$$

Proceeding in this way in can be shown that

$$\psi(t) = 0 \text{ for all} t = 1, 2, \dots N$$

Hence T is complete sufficient for N.

Example 2: Let $X_1, X_2, \ldots, X_n$ be n be a random sample from $R(\theta_1, \theta_2)$. So the density for each of X_i

$$f_{\theta_1,\theta_2}(x) = \frac{1}{(\theta_2 - \theta_1)^n}, \text{ if } \theta_1 \leq x \leq \theta_2.$$

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $X_{(1)} = \min[X_1, X_2, \ldots, X_n]$ and $X_{(n)} = \max[X_1, X_2, \ldots, X_n]$. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is given by

$$f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n) = \frac{1}{\theta_2 - \theta_1}, \text{ if } \theta_1 \leq x_{(1)} < x_{(n)} \leq \theta_2$$

If we define a function $\phi(a, b)$ such that

$$\begin{aligned} \phi(a,b) &= 1, \text{ if } a \leq b \\ &= 0, \text{ if } a > b \end{aligned}$$

then $f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n)$ can be written as

$$f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n) = \frac{1}{\theta_2 - \theta_1}\phi(\theta_1, x_{(1)})\phi(\theta_2, x_{(n)})$$

If we write $t_1 = x_{(1)}$, $t_2 = x_{(n)}$, $g_{\boldsymbol{\theta}}(t_1, t_2) = \frac{1}{\theta_2 - \theta_1}\phi(\theta_1, t_1)\phi(\theta_2, t_2)$ and $h(x_1, x_2, \ldots x_n) = 1$ then $f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n)$ can be factorized as

$$f_{\boldsymbol{\theta}}(x_1, x_2, \ldots x_n) = g_{\boldsymbol{\theta}}(t_1, t_2)h(x_1, x_2, \ldots x_n).$$

Then by factorization theorem $T = (X_{(1)}, X_{(n)})$ is jointly sufficient for $\theta = (\theta_1, \theta_2)$.
We shall prove show that $X_{(1)}$ and $X_{(n)}$ are jointly complete. Now the joint density of $T = (X_{(1)}, X_{(n)})$ is

$$f_{\theta_1,\theta_2}(t_1, t_2) = \frac{n(n-1)}{(\theta_2 - \theta_1)^n}(t_2 - t_1)^{n-2}, \theta_1 \leq t_1 \leq t_2 \leq \theta_2.$$

Let $\psi(T)$ be such that

$$E_\theta(\psi(T)) = \int_{\theta_1}^{\theta_2}\int_{\theta_1}^{t_2}\psi(t_1, t_2)f_\theta(t_1, t_2)dt_1dt_2 = 0, \text{ for all } \theta \in \Theta,$$

i.e.,

$$\int_{\theta_1}^{\theta_2}\int_{\theta_1}^{t_2}\psi(t_1, t_2)(t_2 - t_1)^{n-2} = 0, \text{ for all } \theta_1 < \theta_2.$$

Differentiating both sides with respect to θ_2, by Leibnitz theorem we get

$$\int_{\theta_1}^{\theta_2} \psi(t_1, \theta_2)(t_2 - t_1)^{n-2} = 0, \text{ for all } \theta_1 < \theta_2,$$

Again differentiating with respect to θ_1, we get $\psi(\theta_1, \theta_2)(\theta_2 - \theta_1)^{n-2} = 0$ for all $\theta_1 < \theta_2$. Hence, $\psi(t_1, t_2) = 0$ for $t_1 < t_2$. So, $X_{(1)}, X_{(n)}$ is jointly complete for $U(\theta_1, \theta_2)$.

Some remarks on complete sufficient statistic

Remark 1 : A sufficient statistic may not be complete.

Example : Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $N(\theta, \theta^2), \theta > 0$. Then by factorization theorem, $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is jointly sufficient for θ. If we write $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$ then

$$T_1 \sim N(n\theta, n\theta^2)$$

$$E_\theta(T_1) = n\theta, \quad Var_\theta(T_1) = n\theta^2$$

Hence

$$E_\theta(T_1^2) = n\theta^2 + n^2\theta^2$$

Also $E_\theta(T_2) = \sum_{i=1}^n E_\theta(X_i) = \sum_{i=1}^n [E_\theta^2(X_i + Var_\theta(X_i))] = 2n\theta^2$
If we define a function of (T_1, T_2) as $\psi(T_1, T_2) = 2T_1^2 - (n+1)T_2$ then

$$E_\theta [\psi(T_1, T_2)] = 0 \text{ for all } \theta > 0$$

But $\psi(T_1, T_2) \neq 0$ with probability one. Hence $T = (T_1, T_2)$ is not complete.

Remark 2 : A complete statistic may not be sufficient.

Example : Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random sample from $P(\theta)$ population.
Since $X_1 \sim P(\theta)$. by Result 1, X_1 is a complete statistic but X_1 is not sufficient for θ.

Statistical Inference I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Ancillarity

The concept called *ancillarity* of a statistic is perhaps the furthest away from sufficiency. A sufficient statistic $T(\boldsymbol{X})$ preserves all the information about θ contained in the data $\boldsymbol{X}$.To contrast, an ancillary statistic $V(\boldsymbol{X})$ by itself provides no information about the unknown parameter θ. Individually. an ancillary statistic would not provide any information about θ, but such statistics can play useful roles in statistical methodology. The term ancillary statistic (from the Latin ancilla, meaning handmaiden) was introduced by R. A. Fisher (1925) in the context of maximum likelihood estimation.

Definition A statistic $V(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called ancillary for θ.

Let us consider the following examples

Example 1 Let $X_i \sim R(\theta, \theta+1),\ i = 1, 2 \ldots n$ independently where $-\infty < \theta < \infty$. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics from the sample. The sample range is given by $R = X_{(n)} - X_{(1)}$ is ancillary for θ. To prove this let us define $Z_i = X_i - \theta$, $i = 1, 2 \ldots n$. Then $Z_i \sim R(0, 1)$. $i = 1, 2 \ldots n$ and hence the distribution of Z_i is independent of θ. If $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$ be the order statistics for Z_i's then $R = X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$. Since the distribution of Z_i is independent of θ, the distribution of R is also independent of θ.

Example 2 Let $X_i \sim R(0, \theta),\ i = 1, 2 \ldots n$ independently where $0 < \theta < \infty$. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics from the sample. Then the statistic $V = \frac{X_{(n)}}{X_{(1)}}$ is ancillary for θ. To prove this let us define $Z_i = \frac{X_i}{\theta}$, $i = 1, 2 \ldots n$. Then

$Z_i \sim R(0,1)$. $i = 1,2\ldots n$ and hence the distribution of Z_i is independent of θ.If $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$ be the order statistics for Z_i's then $V = \frac{X_{(n)}}{X_{(1)}} = \frac{Z_{(n)}}{Z_{(1)}}$. Since the distribution of Z_i is independent of θ, the distribution of V is also independent of θ.

Example 3 Let $X_i \sim N(\theta, 1),\ i = 1,2\ldots n$ independently where $-\infty < \theta < \infty$. Then $\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$ and $\sum\limits_{i=1}^{n}|X_i - \bar{X}|$ are ancillary statistics for θ, where $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. To prove this let us define $Z_i = X_i - \theta$ and $W_i = X_i - \bar{X},\ i = 1,2\ldots n$. Then $Z_i \sim N(0,1)$. $i = 1,2\ldots n$ and $\bar{Z} = \bar{X} - \theta \sim N(0, \frac{1}{n})$, both have the distribution independent of θ. As $W_i = Z_i - \bar{Z},\ i = 1,2\ldots n$, the distribution of W_i is also independent of θ. Hence $\sum\limits_{i=1}^{n} W_i^2$ and $\sum\limits_{i=1}^{n}|W_i|$ are ancillary statistics for θ.

Example 4 Let X and Y be independently distributed random variables such that X follows exponential with mean θ and Y follows exponential with mean $\frac{1}{\theta}$ respectively. Then the distributions of $Z = \frac{X}{\theta}$ and $W = \theta Y$ are independent of θ. If we define $V = XY = ZW$, the distribution of V is independent of θ. Thus V is ancillary for θ.

Example 5 Let X denotes the number of points obtained in throwing a biased die and the probability distribution of X be given by

$$P_\theta(X=1) = \frac{1}{12}(1-\theta), P_\theta(X=2) = \frac{1}{12}(2-\theta), P_\theta(X=3) = \frac{1}{12}(3-\theta),$$
$$P_\theta(X=4) = \frac{1}{12}(1+\theta), P_\theta(X=5) = \frac{1}{12}(2+\theta), P_\theta(X=6) = \frac{1}{12}(3+\theta).$$

If we define a random variable V such that

$$V = 0 \ \text{ if } X = 1 \text{ or } 4$$

$$V = 1 \ \text{ if } X = 2 \text{ or } 5$$

$$V = 2 \ \text{ if } X = 3 \text{ or } 6$$

the distribution of V is independent of θ and V is ancillary for θ.

Example 6 Let X be a random variable with p.d.f.

$$f_\theta(x) = 1, \text{ if } \theta \leq x < \theta + 1, \theta \geq 0$$

If we define $V = X - [X]$, where $[X]$ denotes the greatest integer contained X, then V has $R(0,1)$ distribution. Since the the distribution of V is independent of θ the statistic V is ancillary for θ.

Next we consider a very nice theorem which was due to Basu (1955).

Basu's Theorem

If T is complete sufficient for θ and V is ancillary for θ, then V is independent of T.

Proof For simplicity, we prove the theorem only in the discrete case. Suppose that the domain spaces of V and T are respectively denoted by $\mathcal{V}$ and $\mathcal{T}$ respectively.

Let for any $v \in \mathcal{V}$, we write $h(v) = P_\theta(V = v)$. Obviously $h(v)$ is free from θ since V is ancillary for θ.

Since T is sufficient for θ the conditional distribution of V given $T = t$ is independent of θ for all $t \in \mathcal{T}$.

Let us write $g(t) = P(V = v|T = t)$, $t \in \mathcal{T}$ Then

$$\begin{aligned} E_\theta(g(T)) &= \sum_{t \in \mathcal{T}} P(V = v|T = t) P_\theta(T = t) \\ &= \sum_{t \in \mathcal{T}} P_\theta(V = v, T = t) \\ &= P_\theta(V = v) \\ &= h(v) \end{aligned}$$

It follows that $E_\theta(g(T) - h(v)) = 0$ for all $\theta \in \boldsymbol{\theta}$.

Since T is complete we get

$$g(t) = h(v) \quad \text{for all } v \in \mathcal{V}, t \in \mathcal{T}.$$

Hence V and T are independently distributed.

Applications of Basu's theorem

Basu's theorem can be used for proving the independence between two statistics. Some examples are given below.

Application 1 In Example 2, $T = X_{(n)}$ is complete sufficient for θ and it is distributed independently of $V = V = \frac{X_{(n)}}{X_{(1)}}$.

Application 2

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

In this module we are going to discuss to important theorems in statistical inference. The first theorem is on the minimal sufficiency and completeness of a family of distributions and the second theorem is on the improvement of an unbiased estimator through sufficiency.

Theorem I If a minimal sufficient statistic exists for a family $\mathcal{F}_\theta = \{f_\theta(x) : \theta \in \Theta\}$, then a complete sufficient statistic is minimal sufficient.

Proof Let $T^\star$ be a minimal sufficient statistic and T be a complete sufficient statistic for a parameter θ. We have to show that T and T^* are equivalent. As T^* is minimal sufficient statistic it is a function of every other sufficient statistic.

Hence $T^* = h(T)$, a function of T.

Let us now define, $\phi(T) = T - E(T|T^*)$. Since T^* is sufficient, $E(T|T^*)$ is independent of θ and hence $\phi(T)$ is a function of T only.

$$E(\phi(T)) = E\left[T - E(T|T^*)\right] = E(T) - E(T) = 0 \text{ for all } \theta$$

$$\Rightarrow \phi(T) = 0 \text{ a.e. w.r.t.} \mathcal{F}_\theta \text{ as T is complete}$$

$$\Rightarrow T = E(T|T^*) = h^*(T^*) \text{ a.e. w.r.t.} \mathcal{F}_\theta$$

i.e. T is a function of T^* and hence T and T^* are equivalent.

Remark The converse of the theorem may not be true i.e.a minimal sufficient statistic may not be complete.

Counter example Let X be a random variable with p.m.f

$$\begin{aligned} f_\theta(x) &= \theta, \text{ if } x = -1 \\ &= (1-\theta)^2\theta^x, \text{ if } x = 0, 1, 2, \ldots, 0 < \theta < 1. \end{aligned}$$

Clearly X is minimal sufficient for θ. To check the completeness of the family let us consider a function $\psi(X)$ of X such that

$$E_\theta(\psi(X)) = 0 \text{ for all } \theta$$

$$\Rightarrow \theta\psi(-1) + \sum_{x=0}^{\infty} \psi(x)(1-\theta)^2\theta^x = 0 \text{ for all } \theta$$

$$\Rightarrow \sum_{x=0}^{\infty} \psi(x)\theta^x = \frac{-\theta\psi(-1)}{(1-\theta)^2} = -\sum_{x=0}^{\infty} \psi(-1)x\theta^x \text{ for all } \theta$$

Comparing the coefficients of $\theta^x, x = 0, 1, 2, \ldots$ from both sides of the above identity we get

$$\psi(x) = -xh(-1), \text{ for all } x$$

If $\psi(-1) = c \neq 0$, then $\psi(x) = -cx$, for all x, i.e. $\psi(x)$ is non-zero with non-zero probability.
Hence the family of distributions $\mathcal{F}_\theta = \{f_\theta(x) : \boldsymbol{\theta} \in \Theta\}$ is not complete.

Theorem II Let the statistic $U = U(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of $g(\theta)$ while T is a sufficient statistic for θ. Consider the function $\phi(T)$ of T such that,

$$\phi(t) = E(U|T = t).$$

Then $\phi(T)$ is itself a unbiased estimator of $g(\theta)$; moreover,

$$Var_\theta(\phi(T)) \leq Var_\theta(U)$$

Proof Since T is sufficient for θ, the conditional expectation of U given $T = t$ can not depend upon the unknown parameter θ. Hence $\phi(t)$ is a function of t and it is free from θ. In other words, $\phi(T)$ is a real valued statistic and so we can call it an estimator. Using a theorem on conditional expectation (see Rohatgi and Saleh) we get

$$E_\theta(U) = E_\theta(E(U|T)) = E_\theta(\phi(T)),$$

i.e.

$$E_\theta(\phi(T)) = g(\theta), \qquad \forall\theta \in \Theta.$$

Hence $\phi(T)$ is an unbiased estimator of $g(\theta)$. By a theorem on conditional variance (see Rohatgi and Saleh) we get

$$\begin{aligned} Var_\theta(U) &= E_\theta[Var(U|T)] + Var_\theta[E(U|T)] \\ &= E_\theta[Var(U|T)] + Var_\theta[\phi(T)]. \end{aligned}$$

Since $E_\theta[Var(U|T)] \geq 0$ we have

$$Var_\theta(U) \geq Var_\theta[\phi(T)], \text{ for all } \theta \in \Theta$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

Equality sign holds if

$$E_\theta[Var(U|T)] = 0, \text{ for all } \theta$$

$$\Rightarrow E_\theta[E\{U - E(U|T)\}^2|T] = 0, \text{ for all } \theta$$
$$\Rightarrow E_\theta\{U - \phi(T)\}^2 = 0, \text{ for all } \theta$$
$$\Rightarrow P_\theta(U = \phi(T)) = 1, \text{ for all } \theta.$$

Note In literature the above theorem is known as **Rao-Blackwell** theorem. The technique to improve upon an initial unbiased estimator of $g(\theta)$ is customarily referred to as the Rao-Blackwellization in the statistical literature. One attractive feature of the Rao-Blackwell Theorem is that there is no need to guess the functional form of the final unbiased estimator of $g(\theta)$.

Implication Since an unbiased estimator based on a sufficient statistic T will always has a smaller variance than one which is not based on T, the search for an UMVUE may be restricted to those unbiased estimators which are based on the sufficient statistic T.

Some examples:

Example 1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a $Bin(1, \theta)$ population with p.m.f.

$$f_\theta(x) = \theta^x(1-\theta)^{1-x}, x = 0, 1.$$

If we consider the function $g(\theta) = \theta$ then X_1 is an unbiased estimator of $g(\theta)$. Our object is to find the improved unbiased estimator for $\gamma(\theta) = \theta$ based on the sufficient statistic. Note that, $T = \sum_{i=1}^n X_i$ is sufficient for θ. Consider the statistic $\phi(T)$ such that

$$\phi(t) = E(X_1|T = t) = P\left[X_1 = 1|\sum_{i=1}^n X_i = t\right]$$

Then by definition of conditional probability we get

$$\begin{aligned}
\phi(t) &= \frac{P\left[X_1 = 1, \sum_{i=1}^n X_i = t\right]}{P\left[\sum_{i=1}^n X_i = t\right]} \\
&= \frac{P\left[X_1 = 1, \sum_{i=2}^n X_i = t-1\right]}{P\left[\sum_{i=1}^n X_i = t\right]} \\
&= \frac{P[X_1 = 1]P\left[\sum_{i=2}^n X_i = t-1\right]}{P\left[\sum_{i=1}^n X_i = t\right]}
\end{aligned}$$

$$= \frac{(1-\theta)\binom{(n-1)}{(t-1)}\theta^{t-1}(1-\theta)^{n-1-t+1}}{\binom{n}{s}\theta^t(1-\theta)^{n-t}}$$
$$= \frac{t}{n}$$

Therefore, $\phi(T) = \frac{1}{n}\sum_{i=1}^{n} X_i$ Rao-Blackwellized version of the initial unbiased estimator X_1.

Example 2 Let $X_1, X_2, \ldots, X_n$ be a random sample from a $Poisson(\theta)$ with p.m.f.

$$\begin{aligned} f_\theta(x) &= \frac{\exp(-\theta)\theta^x}{x!} \text{ if } x = 0, 1, \ldots, \\ &= 0 \text{ otherwise} \end{aligned}$$

Consider the function

$$g(\theta) = P[X = k] = \frac{\exp(-\theta)\theta^k}{k!}.$$

Let us define a random variable U such that,

$$\begin{aligned} U &= 1 \text{ if } X_1 = k \\ &= 0 \text{ otherwise} \end{aligned}$$

Then

$$\begin{aligned} E_\theta(U) &= 1 \times P_\theta[X = k] + 0 \times P[X \neq k] \\ &= g(\theta) \text{ for all } \theta > 0, \end{aligned} \tag{1}$$

so that U is an unbiased of $g(\theta)$. Now we know that $T = \sum_i X_i$ is a sufficient statistic for θ. Consider then the new statistic $\phi(T)$ such that,

$$\begin{aligned} \phi(t) = E(Y|T = t) &= P[X_1 = k|T = t] \\ &= \frac{P[X_1 = k, \sum_{i=1}^{n} X_i]}{P[T = t]} \\ &= \frac{\exp(-\theta)\theta^k}{k!} . \frac{\exp(-(n-1)\theta)[(n-1)\theta]^{t-k}}{(t-k)!} \Big/ \frac{\exp[-n\theta](n\theta)^t}{t!} \\ &= \binom{n}{k}\frac{(n-1)^{t-k}}{n^t}. \end{aligned}$$

since T has the Poisson distribution with parameter $n\theta$ while $\sum_{i=1}^{n} X_i$ has the Poisson distribution with parameter $(n-1)\theta$. Thus $\phi(T) = \binom{T}{k}\frac{(n-1)^{T-k}}{n^T}$ is an improved unbiased estimator of $g(\theta)$ with a smaller variance of the initial unbiased estimator U of $g(\theta)$.

Example 3 Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ with p.d.f.

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right] - \infty < x < \infty.$$

If we consider the function $g(\theta) = \theta$ then X_1 is an unbiased estimator of $g(\theta)$. Our object is to find the improved unbiased estimator for $\gamma(\theta) = \theta$ based on the sufficient statistic. Note that, $T = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sufficient for θ. Consider the statistic $\phi(T)$ such that

$$\phi(t) = E(X_1|T = t)$$

Since the conditional distribution of X_1 given $T = t$ is $N(t, 1 - \frac{1}{n})$ it follows that $E(X_1|T = t) = t$. Hence Rao-Blackwellized version of X_1 is T.

Example 4 Let $X_1, X_2, \ldots, X_n$ be a random sample from a $Rect(0, \theta)$ distribution with p.d.f.

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta}, 0 < x < \theta \\ &= 0 \text{ otherwise} \end{aligned}$$

Consider the function

$$g(\theta) = \frac{\theta}{2}$$

Since $E_\theta(X_1) = \frac{\theta}{2}$ for all $\theta > 0$, we consider $U = X_1$ as an initial unbiased estimator of $g(\theta)$. Here the statistic $T = Max(X_1, X_2, \ldots, X_n)$ is sufficient for θ. To improve our initial estimator let us define the statistic

$$\phi(T) = E(U|T).$$

To find $\phi(T)$ note that if $T = t$, then $X_1 = t$ with probability $\frac{1}{n}$, and X_1 is uniformly distributed over $(0, t)$ with remaining probability $\left(1 - \frac{1}{n}\right)$.Hence

$$E(X_1|T = t) = t\frac{1}{n} + \frac{t}{2}\left(1 - \frac{1}{n}\right) = \frac{n+1}{n}\frac{t}{2}.$$

Thus $\phi(T) = \frac{n+1}{n}\frac{T}{2}$ is Rao-Blackwellized version of the initial unbiased estimator U.

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Subject-Statistics
Paper- Determination of UMVUE by Complete Sufficient Statistic

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

If the Rao-Blackwellized version in the end always comes up with the same refined estimator regardless of which unbiased estimator of $g(\theta)$ one initially starts with, then of course one has found the UMVUE for $g(\theta)$. Lehmann and Scheffe.s (1950) notion of a complete statistic, introduced in Module-9, plays a major role in this area. We first prove the following result from Lehmann and Scheffe (1950).

Lehmann-Scheffe Theorem I Suppose that U is an unbiased estimator of a real valued parametric function $g(\theta)$ where the unknown parameter $\theta \in \Theta \subseteq \mathcal{R}$. Let T be a complete sufficient statistic for θ. If we define a statistic $\phi(T) = E(U|T)$, then $\phi(T)$ is the unique (i.e. w.p. 1) of $g(\theta)$.
Note: In this theorem the parameter θ may be vector valued and so is the complete sufficient statistic T.
Proof The Rao-Blackwell theorem assures us that in order to search for the UMVUE of $g(\theta)$ we need only to focus on unbiased estimators which are functions of sufficient statistic T alone. We already know that $\phi(T)$ is a function of T and and it is an unbiased estimator of $g(\theta)$. Suppose that there is another unbised estimator $\phi^*(T)$ of $g(\theta)$ where $\phi^*(T)$ is also a function of T.
If we define $h(T) = \phi(T) - \phi^*(T)$, then

$$E_\theta(h(T)) = g(\theta) - g(\theta) = 0 \text{ for all } \theta \tag{1}$$

Now, we use the definition of the completeness of a statistic. Since T is a complete statistic, from (1) it follows that $h(T) = 0$ w.p. 1, that is we must have $\phi(T) = \phi^*(T)$ w.p. 1. The result then follows.

Note : In order to find the UMVUE of $g(\theta)$, we need not always have to go through the conditioning with respect to the complete sufficient statistic T. In some problems, the following alternate and yet equivalent result may be more directly applicable. We state the result without proving it. Its proof

can be easily constructed from the proof of Lehmann-Scheffe Theorem I.

Lehmann-Scheffe Theorem II Suppose that T be a complete sufficient statistic for $\theta in \Theta$. If $\psi(T)$ is a unbiased estimator of the of a real valued parametric function $g(\theta)$, then $\psi(T)$ is the unique (i.e. w.p. 1) of $g(\theta)$.

Let us consider some examples.
Example 1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with probability mass function,

$$\begin{aligned} f_\theta(x) &= \frac{\exp(-\theta)\theta^x}{x!} \text{ if } x = 0, 1, \ldots, \\ &= 0 \text{ otherwise} \end{aligned}$$

We are interested in finding the UMVUE of $g(\theta) = e^{-\theta} = P_\theta(X_1 = 0)$. We know that, $f_\theta(x)$ belongs to the exponential family. Therefore $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic. Consider the following statistic,

$$\begin{aligned} U &= 1 \text{ if } X_1 = 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

Then we get, $E_\theta[U] = P_\theta[X_1 = 0] = \exp(-\theta)$. Hence U is an unbiased estimator of $g(\theta)$. If we define $\phi(T) = E(U|T)$ then,

$$\begin{aligned} \phi(t) = P\left[X_1 = 0 | T = t\right] &= \frac{P\left[X_1 = 0, \sum_{i=1}^n X_i = t\right]}{P\left[\sum_{i=1}^n X_i = t\right]} \\ &= \frac{P\left[X_1 = 0, \sum_{i=2}^n X_i = t\right]}{P\left[\sum_{i=1}^n X_i = t\right]} \\ &= \frac{P[X_1 = 0] P\left[\sum_{i=2}^n X_i = t\right]}{P\left[\sum_{i=1}^n X_i = t\right]} \\ &= \frac{\exp(-\theta)\exp(-(n-1)\theta)[(n-1)\theta]^t/t!}{\exp(-(n)\theta)[(n)\theta]^t/t!} \\ &= \left(\frac{n-1}{n}\right)^t. \end{aligned}$$

Hence,by Lehmann-Scheffe theorem, $\phi(T) = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}$ is the UMVUE of $e^{-\theta}$ for $n > 1$.

Example 2: Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where $\mu \in \mathcal{R}$ and $\sigma \in \mathcal{R}^+$ are unknown. Suppose we want to find the UMVUE of μ^2. If we define $\boldsymbol{\theta} = (\mu, \sigma^2)$ then $\boldsymbol{\theta} \in \Theta = \mathcal{R} \times \mathcal{R}^+$. Here the statistic $T = (\bar{X}, S^2)$ is complete sufficient for $\boldsymbol{\theta}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, we have $E_\theta(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$.
As $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, we have $E_\theta(S^2) = \sigma^2$. Hence we get $E_\theta(\bar{X}^2 - \frac{S^2}{n}) = \mu^2$, for all θ Since $\bar{X}^2 - \frac{S^2}{n}$ is an UE of μ^2 based on the complete sufficient T, by Lehmann-Scheffe theorem $\bar{X}^2 - \frac{S^2}{n}$ is the UMVUE of μ^2.

Example 3: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Rectangular distribution $R(0, \theta)$ where $0 < \theta < \infty$ is unknown. Our problem is to find the UMVUE for θ^α where α is a known positive real number.
The complete sufficient statistic for θ is $T = X_{(n)} = Max(X_1, X_2, \ldots, X_n)$. The p.d.f. of $X_{(n)}$ is given by

$$f_\theta(t) = \frac{n}{\theta^n} t_{(n)}^{n-1}, 0 < t < \theta.$$

Now,

$$E_\theta(T^\alpha) = \frac{n}{\theta^n}\int_0^\theta t^{\alpha+n-1} dt = \frac{n}{n+\alpha}\theta^\alpha, \text{ for all} \theta > 0,$$
$$\Rightarrow E_\theta\left(\frac{n+\alpha}{n}T^\alpha\right) = \theta^\alpha, \text{ for all } \theta > 0$$

Hence $\frac{n+\alpha}{n}T^\alpha$ is an UE of θ^α based on a complete sufficient statistic T. Hence,by Lehmann-Scheffe theorem, it is the UMVUE of θ^α.

Example 4: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Rectangular distribution $R(\theta_1, \theta_2)$ where $0 < \theta_1 < \theta_2 < \infty$ are unknown. Our problem is to find the UMVUE for $\theta_2 - \theta_1$ and $\theta_2 + \theta_1$. Note that the joint density is given by,

$$f_{\theta_1,\theta_2}(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n \text{ if, } \theta_1 < x_1, x_2, \ldots, x_n < \theta_2.$$

(i). We know that $T = (X_{(1)}, X_{(n)})$ are jointly complete sufficient statistic for θ_1, θ_2, where $X_{(1)}$ is the smallest order statistic and $X_{(n)}$ is the largest order

ePathshala पाठशाला A Gateway to All Post Graduate Courses

statistic.
The p.d.f. of $X_{(n)}$ is given by

$$f_{\theta_1,\theta_2}(x_{(n)}) = \frac{n}{(\theta_2-\theta_1)^n}(x_{(n)}-\theta_1)^{n-1} \text{ if, } \theta_1 < x_{(n)} < \theta_2,$$

and hence,

$$E(X_{(n)}-\theta_1) = \int_{\theta_1}^{\theta_2} \frac{n(x_{(n)}-\theta_1)^n}{(\theta_2-\theta_1)^n} d(x_{(n)}-\theta_1) = \frac{n}{n+1}(\theta_2-\theta_1) \qquad (1)$$

. The p.d.f. of $X_{(1)}$ is given by

$$f_{\theta_1,\theta_2}(x_{(1)}) = \frac{n}{(\theta_2-\theta_1)^n}(\theta_2-x_{(1)})^{n-1} \text{ if, } \theta_1 < x_{(1)} < \theta_2,$$

and hence,

$$E(\theta_2-X_{(1)}) = \int_{\theta_1}^{\theta_2} \frac{n(\theta_2-x_{(1)})^n}{(\theta_2-\theta_1)^n} d(\theta_2-x_{(1)}) = \frac{n}{n+1}(\theta_2-\theta_1) \qquad (2)$$

. Therefore, adding (1) and (2) and we get

$$\begin{aligned} E(X_{(n)}-X_{(1)}+(\theta_2-\theta_1)) &= \frac{2n}{n+1}(\theta_2-\theta_1) \\ \Rightarrow E(X_{(n)}-X_{(1)}) &= \frac{n-1}{n+1}(\theta_2-\theta_1). \\ \Rightarrow E\left[\frac{n-1}{n+1}(X_{(n)}-X_{(1)})\right] &= (\theta_2-\theta_1). \end{aligned}$$

Again, subtracting (2) from (1) and we get

$$\begin{aligned} E(X_{(n)}+X_{(1)}-(\theta_1+\theta_2)) &= 0 \\ \Rightarrow E(X_{(n)}+X_{(1)}) &= \theta_1+\theta_2. \end{aligned}$$

Hence, $\frac{n+1}{n-1}\left(X_{(n)}-X_{(1)}\right)$ and $X_{(n)}+X_{(1)}$ are the unbiased estimators of $\theta_2-\theta_1$ and $\theta_2+\theta_1$ based on complete sufficient statistic $T=(X_{(1)},X_{(n)})$. Hence,by Lehmann-Scheffe's theorem, $\frac{n+1}{n-1}\left(X_{(n)}-X_{(1)}\right)$ and $X_{(n)}+X_{(1)}$ are the uniformly minimum variance unbiased estimators of $\theta_2-\theta_1$ and $\theta_2+\theta_1$ respectively.

A Gateway to All Post Graduate Courses

Example 5: Let $X_1, X_2, \ldots, X_n$ be a sample of discrete random variables from $N(\mu, \sigma^2)$, where $\mu \in \mathcal{R}$ is unknown but $\sigma > 0$ is unknown. We wish to find the UMVUE of $\pi_c = P[X_1 \leq c]$, . The complete sufficient statistic for μ is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We start we the statistic

$$\begin{aligned} U &= 1 \text{ if } X_1 \leq c \\ &= 0 \text{ otherwise} \end{aligned}$$

Then we get, $E_\mu(U) = P[X_1 \leq c] = \pi_c$ for all μ. Therefore, U is an unbiased estimator of π_c. If we define $\phi(\bar{X}) = E(U|\bar{X})$, then

$$\begin{aligned} \phi(\bar{X}) &= P(X_1 \leq c|\bar{X}) \\ &= P(X_1 - \bar{X} \leq c - \bar{X}|\bar{X}), \end{aligned}$$

Now $(X_1 - \bar{X}) \sim N(0, \frac{n-1}{n})$, which is independent of μ. Since $(X_1 - \bar{X})$ is ancillary for μ and $\bar{X}$ complete sufficient for μ, by Basu's theorem they are independently distributed. Hence

$$\phi(\bar{X}) = P(X_1 - \bar{X} \leq c - \bar{X}) = \Phi\left(\sqrt{\frac{n}{n-1}}(c - \bar{X})\right),$$

where $\Phi(.)$ is the distribution function of standard normal distribution. Therefore $\phi(\bar{X})$ is an UE of π_c based on complete sufficient statistic $\bar{X}$. Hence by Lehmann-Scheffe's theorem, $\Phi\left(\sqrt{\frac{n}{n-1}}(c - \bar{X})\right)$ is the UMVUE of π_c.

Pathshala A Gateway to All Post Graduate Courses

Subject-Statistics
Paper- Determination of UMVUE by Complete Sufficient Statistic
Module- 14

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

If the Rao-Blackwellized version in the end always comes up with the same refined estimator regardless of which unbiased estimator of $g(\theta)$ one initially starts with, then of course one has found the UMVUE for $g(\theta)$. Lehmann and Scheffe.s (1950) notion of a complete statistic, introduced in Module-9, plays a major role in this area. We first prove the following result from Lehmann and Scheffe (1950).

Lehmann-Scheffe Theorem I Suppose that U is an unbiased estimator of a real valued parametric function $g(\theta)$ where the unknown parameter $\theta \in \Theta \subseteq \mathcal{R}$. Let T be a complete sufficient statistic for θ. If we define a statistic $\phi(T) = E(U|T)$, then $\phi(T)$ is the unique (i.e. w.p. 1) UMVUE of $g(\theta)$.

Note: In this theorem the parameter θ may be vector valued and so is the complete sufficient statistic T.

Proof The Rao-Blackwell theorem assures us that in order to search for the UMVUE of $g(\theta)$ we need only to focus on unbiased estimators which are functions of sufficient statistic T alone. We already know that $\phi(T)$ is a function of T and and it is an unbiased estimator of $g(\theta)$. Suppose that there is another unbised estimator $\phi^*(T)$ of $g(\theta)$ where $\phi^*(T)$ is also a function of T.

If we define $h(T) = \phi(T) - \phi^*(T)$, then

$$E_\theta(h(T)) = g(\theta) - g(\theta) = 0 \text{ for all } \theta \tag{1}$$

Now, we use the definition of the completeness of a statistic. Since T is a complete statistic, from (1) it follows that $h(T) = 0$ w.p. 1, that is we must have $\phi(T) = \phi^*(T)$

w.p. 1. The result then follows.

Note : In order to find the UMVUE of $g(\theta)$, we need not always have to go through the conditioning with respect to the complete sufficient statistic T. In some problems, the following alternate and yet equivalent result may be more directly applicable. We state the result without proving it. Its proof can be easily constructed from the proof of Lehmann-Scheffe Theorem I.

Lehmann-Scheffe Theorem II Suppose that T be a complete sufficient statistic for $\theta in \Theta$. If $\psi(T)$ is a unbiased estimator of the of a real valued parametric function $g(\theta)$, then $\psi(T)$ is the unique (i.e. w.p. 1) UMVUE of $g(\theta)$.

Let us consider some examples.

Example 1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with probability mass function,

$$\begin{aligned} f_\theta(x) &= \frac{\exp(-\theta)\theta^x}{x!} \text{ if } x = 0, 1, \ldots, \\ &= 0 \text{ otherwise} \end{aligned}$$

We are interested in finding the UMVUE of $g(\theta) = e^{-\theta} = P_\theta(X_1 = 0)$. We know that, $f_\theta(x)$ belongs to the exponential family. Therefore $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic. Consider the following statistic,

$$\begin{aligned} U &= 1 \text{ if } X_1 = 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

Then we get, $E_\theta[U] = P_\theta[X_1 = 0] = \exp(-\theta)$. Hence U is an unbiased estimator of $g(\theta)$. If we define $\phi(T) = E(U|T)$ then,

$$\phi(t) = P\left[X_1 = 0 | T = t\right] = \frac{P\left[X_1 = 0, \sum_{i=1}^n X_i = t\right]}{P\left[\sum_{i=1}^n X_i = t\right]}$$

$$
\begin{aligned}
&= \frac{P\left[X_1 = 0, \sum_{i=2}^{n} X_i = t\right]}{P\left[\sum_{i=1}^{n} X_i = t\right]} \\
&= \frac{P[X_1 = 0]P\left[\sum_{i=2}^{n} X_i = t\right]}{P\left[\sum_{i=1}^{n} X_i = t\right]} \\
&= \frac{\exp(-\theta)\exp(-(n-1)\theta)[(n-1)\theta]^t/t!}{\exp(-(n)\theta)[(n)\theta]^t/t!} \\
&= \left(\frac{n-1}{n}\right)^t .
\end{aligned}
$$

Hence,by Lehmann-Scheffe theorem, $\phi(T) = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}$ is the UMVUE of $e^{-\theta}$ for $n > 1$.

Example 2: Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where $\mu \in \mathcal{R}$ and $\sigma \in \mathcal{R}^+$ are unknown. Suppose we want to find the UMVUE of μ^2. If we define $\boldsymbol{\theta} = (\mu, \sigma^2)$ then $\boldsymbol{\theta} \in \Theta = \mathcal{R} \times \mathcal{R}^+$. Here the statistic $T = (\bar{X}, S^2)$ is complete sufficient for $\boldsymbol{\theta}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, we have $E_\theta(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$.
As $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, we have $E_\theta(S^2) = \sigma^2$. Hence we get $E_\theta(\bar{X}^2 - \frac{S^2}{n}) = \mu^2,$ for all θ
Since $\bar{X}^2 - \frac{S^2}{n}$ is an UE of μ^2 based on the complete sufficient T, by Lehmann-Scheffe theorem $\bar{X}^2 - \frac{S^2}{n}$ is the UMVUE of μ^2.

Example 3: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Rectangular distribution $R(0, \theta)$ where $0 < \theta < \infty$ is unknown. Our problem is to find the UMVUE for θ^α where α is a known positive real number.
The complete sufficient statistic for θ is $T = X_{(n)} = Max(X_1, X_2, \ldots, X_n)$. The p.d.f. of $X_{(n)}$ is given by

$$f_\theta(t) = \frac{n}{\theta^n} t_{(n)}^{n-1}, 0 < t < \theta.$$

Now,

$$E_\theta(T^\alpha) = \frac{n}{\theta^n}\int_0^\theta t^{\alpha+n-1} dt = \frac{n}{n+\alpha}\theta^\alpha, \text{ for all} \theta > 0,$$

$$\Rightarrow E_\theta\left(\frac{n+\alpha}{n} T^\alpha\right) = \theta^\alpha, \text{ for all } \theta > 0$$

Hence $\frac{n+\alpha}{n}T^{\alpha}$ is an UE of θ^{α} based on a complete sufficient statistic T. Hence,by Lehmann-Scheffe theorem, it is the UMVUE of θ^{α}.

Example 4: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Rectangular distribution $R(\theta_1, \theta_2)$ where $0 < \theta_1 < \theta_2 < \infty$ are unknown. Our problem is to find the UMVUE for $\theta_2 - \theta_1$ and $\theta_2 + \theta_1$. Note that the joint density is given by,

$$f_{\theta_1,\theta_2}(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n \text{ if, } \theta_1 < x_1, x_2, \ldots, x_n < \theta_2.$$

(i). We know that $T = (X_{(1)}, X_{(n)})$ are jointly complete sufficient statistic for θ_1, θ_2, where $X_{(1)}$ is the smallest order statistic and $X_{(n)}$ is the largest order statistic.
The p.d.f. of $X_{(n)}$ is given by

$$f_{\theta_1,\theta_2}(x_{(n)}) = \frac{n}{(\theta_2 - \theta_1)^n}(x_{(n)} - \theta_1)^{n-1} \text{ if, } \theta_1 < x_{(n)} < \theta_2,$$

and hence,

$$E(X_{(n)} - \theta_1) = \int_{\theta_1}^{\theta_2} \frac{n(x_{(n)} - \theta_1)^n}{(\theta_2 - \theta_1)^n} d(x_{(n)} - \theta_1) = \frac{n}{n+1}(\theta_2 - \theta_1) \tag{1}$$

. The p.d.f. of $X_{(1)}$ is given by

$$f_{\theta_1,\theta_2}(x_{(1)}) = \frac{n}{(\theta_2 - \theta_1)^n}(\theta_2 - x_{(1)})^{n-1} \text{ if, } \theta_1 < x_{(1)} < \theta_2,$$

and hence,

$$E(\theta_2 - X_{(1)}) = \int_{\theta_1}^{\theta_2} \frac{n(\theta_2 - x_{(1)})^n}{(\theta_2 - \theta_1)^n} d(\theta_2 - x_{(1)}) = \frac{n}{n+1}(\theta_2 - \theta_1) \tag{2}$$

. Therefore, adding (1) and (2) and we get

$$\begin{aligned} E(X_{(n)} - X_{(1)} + (\theta_2 - \theta_1)) &= \frac{2n}{n+1}(\theta_2 - \theta_1) \\ \Rightarrow E(X_{(n)} - X_{(1)}) &= \frac{n-1}{n+1}(\theta_2 - \theta_1). \\ \Rightarrow E\left[\frac{n-1}{n+1}(X_{(n)} - X_{(1)})\right] &= (\theta_2 - \theta_1). \end{aligned}$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

Again, subtracting (2) from (1) and we get

$$\begin{aligned} E(X_{(n)} + X_{(1)} - (\theta_1 + \theta_2)) &= 0 \\ \Rightarrow E(X_{(n)} + X_{(1)}) &= \theta_1 + \theta_2. \end{aligned}$$

Hence, $\frac{n+1}{n-1}\left(X_{(n)} - X_{(1)}\right)$ and $X_{(n)} + X_{(1)}$ are the unbiased estimators of $\theta_2 - \theta_1$ and $\theta_2 + \theta_1$ based on complete sufficient statistic $T = (X_{(1)}, X_{(n)})$. Hence,by Lehmann-Scheffe's theorem, $\frac{n+1}{n-1}\left(X_{(n)} - X_{(1)}\right)$ and $X_{(n)} + X_{(1)}$ are the uniformly minimum variance unbiased estimators of $\theta_2 - \theta_1$ and $\theta_2 + \theta_1$ respectively.

Example 5: Let $X_1, X_2, \ldots, X_n$ be a sample of discrete random variables from $N(\mu, \sigma^2)$, where $\mu \in \mathcal{R}$ is unknown but $\sigma > 0$ is unknown. We wish to find the UMVUE of $\pi_c = P[X_1 \leq c]$, . The complete sufficient statistic for μ is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We start we the statistic

$$\begin{aligned} U &= 1 \text{ if } X_1 \leq c \\ &= 0 \text{ otherwise} \end{aligned}$$

Then we get, $E_\mu(U) = P[X_1 \leq c] = \pi_c$ for all μ. Therefore, U is an unbiased estimator of π_c. If we define $\phi(\bar{X}) = E(U|\bar{X})$, then

$$\begin{aligned} \phi(\bar{X}) &= P(X_1 \leq c|\bar{X}) \\ &= P(X_1 - \bar{X} \leq c - \bar{X}|\bar{X}), \end{aligned}$$

Now $(X_1 - \bar{X}) \sim N(0, \frac{n-1}{n})$, which is independent of μ. Since $(X_1 - \bar{X})$ is ancillary for μ and $\bar{X}$ complete sufficient for μ, by Basu's theorem they are independently distributed. Hence

$$\phi(\bar{X}) = P(X_1 - \bar{X} \leq c - \bar{X}) = \Phi\left(\sqrt{\frac{n}{n-1}}(c - \bar{X})\right),$$

where $\Phi(.)$ is the distribution function of standard normal distribution. Therefore $\phi(\bar{X})$ is an UE of π_c based on complete sufficient statistic $\bar{X}$. Hence by Lehmann-Scheffe's theorem, $\Phi\left(\sqrt{\frac{n}{n-1}}(c - \bar{X})\right)$ is the UMVUE of π_c.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Subject-Statistics
Paper- Cramer-Rao lower bound

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Cramer-Rao lower bound of the variance of an unbiased estimator
In this section we consider a very important inequality which provides a lower bound for the variance of an unbiased estimator. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from $f_\theta(\cdot)$, where θ belongs to Θ. Let $T = t(\mathbf{X})$ be an unbiased estimator of $g(\theta)$. We will consider the case where $f_\theta(\cdot)$ is a probability density function; the development for discrete density functions is analogous. Suppose $\mathcal{F}_\theta = \{f_\theta(\mathbf{x}) : \theta \in \mathbf{\Theta}\}$ be the family of joint p.d.f.'s of $\mathbf{X}$. It satisfies the Cramer-Rao regularity conditions:

1. Θ is an open sub interval of the real line.
2. The support $\mathcal{X} = \{\mathbf{x} : f_\theta(\mathbf{x}) > 0\}$ does not depend on θ.
3. $\frac{\partial}{\partial\theta} \log f_\theta(\mathbf{x})$ exists for all $\mathbf{x}$ and for all θ.
4. For any statistic $h(\mathbf{X})$ with $E_\theta(|h(\mathbf{X})|) < \infty$ for all θ, the the operations of integration and differentiation with respect to θ can be interchanged in $E_\theta(h(\mathbf{X}))$. That is, $\frac{\partial}{\partial\theta} E_\theta(h(\mathbf{X})) = \frac{\partial}{\partial\theta} \int h(\mathbf{x}) f_\theta(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) \frac{\partial}{\partial\theta} f_\theta(\mathbf{x}) d\mathbf{x}$ for all θ.
5. Fisher's information $I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(\mathbf{x})\right)^2\right]$ exists and is positive for all θ.

Theorem Let $\mathcal{F}_\theta$ satisfies the above regularity conditions and T be an UE of an estimable parametric function $g(\theta)$ such that $g'(\theta) = \frac{d}{d\theta} g(\theta)$ exists for all θ. If $Var_\theta(T)$ finite then

$$Var_\theta(T) \geq \frac{[g'(\theta)]^2}{I(\theta)}, \forall\theta$$

Note : The quantity in the right-hand side of the above inequality is called the Cramer-Rao lower bound (CRLB) for the variance of an unbiased estimator of $g(\theta)$.

Proof Let us define the score function $S(\mathbf{x},\theta) = \frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x}) = \frac{1}{f_\theta(\mathbf{x})}\frac{\partial}{\partial\theta}f_\theta(\mathbf{x})$. Then we get,

$$\begin{aligned} E_\theta(S) &= \int \frac{1}{f_\theta(\mathbf{x})}\frac{\partial}{\partial\theta}f_\theta(\mathbf{x})f_\theta(\mathbf{x})d\mathbf{x} \\ &= \int \frac{\partial}{\partial\theta}f_\theta(\mathbf{x})d\mathbf{x} \\ &= \frac{\partial}{\partial\theta}\int f_\theta(\mathbf{x})d\mathbf{x} \ \ \forall\theta \\ &= \frac{\partial}{\partial\theta}1 = 0. \end{aligned}$$

$Var_\theta(S) = E_\theta(S^2) = E_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right)^2$.

$$\begin{aligned} Cov_\theta(S,T) &= E_\theta(TS) \\ &= \int t(\mathbf{x})\frac{1}{f_\theta(\mathbf{x})}\frac{\partial}{\partial\theta}f_\theta(\mathbf{x})f_\theta(\mathbf{x})d\mathbf{x} \\ &= \frac{\partial}{\partial\theta}\int t(\mathbf{x})f_\theta(\mathbf{x})d\mathbf{x} \\ &= \frac{\partial}{\partial\theta}E_\theta(T(\mathbf{x})) = \frac{\partial}{\partial\theta}g(\theta) = g'(\theta). \end{aligned}$$

If $\rho_\theta(T,S)$ is the correlation between T and S, then

$$\begin{aligned} \rho^2_\theta(T,S) &\leq 1 \\ \Rightarrow Cov^2_\theta(T,S) &\leq Var_\theta(S)Var_\theta(T) \\ \Rightarrow (g'(\theta))^2 &= I(\theta)Var_\theta(T) \\ \Rightarrow Var_\theta(T) &\leq \frac{(g'(\theta))^2}{I(\theta)}. \end{aligned}$$

Equality sign holds if $\rho_\theta(T,S) = 1$ which implies that T is a linear function of S with a positive slope i.e.

$$T = a(\theta) + b(\theta)S \text{ with probability } 1 \qquad (1)$$

where $a(\theta)$ and $b(\theta)$ are constants with $b(\theta) > 0$. Since $E_\theta(S) = 0$, taking expectations in (1) we get, $E_\theta(T) = a(\theta)$ which implies that $a(\theta) = g(\theta)$. Since $V_\theta(S) = I(\theta)$, from (1) we get, $b(\theta) = \frac{g'(\theta)}{I(\theta)}$.Therefore to hold the equality in CRLB we must have

$$T - g(\theta) = \frac{g'(\theta)}{I(\theta)}\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right] \text{ with probability 1} \qquad (2).$$

epathshala A Gateway to All Post Graduate Courses

Note : If a family of distributions satisfies the Cramer-Rao regularity conditions and if $T(\mathbf{X})$ be an UE of an estimable and differentiable parametric function $g(\theta)$ such that (2) holds, then $T(\mathbf{X})$ is the UMVUE of $g(\theta)$.
Now we consider the following results.
Result 1 In Cramer-Rao inequality, equality hold if and only if $f_\theta(x)$ is of exponential form.
Proof. *If part,* Given that the family, is of exponential form,

$$\begin{aligned} f_\theta(x) &= u(\theta)\exp^{q(\theta)t(x)} v(x) \\ &= \exp^{\log u(\theta)+q(\theta)t(x)+\log v(x)} \\ \log f_\theta(x) &= \log u(\theta) + q(\theta)t(x) + \log v(x) \\ &= k(\theta) + q(\theta)t(x) + h(x), \end{aligned}$$

where, $k(\theta) = \log u(\theta)$ and $h(x) = \log v(x)$. Therefore,

$$\begin{aligned} \frac{\partial}{\partial\theta}\log f_\theta(x) &= k'(\theta) + q'(\theta)t(x) \\ t(x) &= \frac{1}{q'(\theta)}\frac{\partial}{\partial\theta}\log f_\theta(x) - \frac{k'(\theta)}{q'(\theta)} \\ t(x) - \left(-\frac{k'(\theta)}{q'(\theta)}\right) &= \frac{1}{q'(\theta)}\frac{\partial}{\partial\theta}\log f_\theta(x), \end{aligned}$$

and hence the equality condition for Cramer-Rao lower bound is attained.
Only if part. Given that the equality condition in Cramer-Rao lower bound holds,

$$\begin{aligned} T(x) &= g(\theta) + \frac{g'(\theta)}{I(\theta)}\frac{\partial}{\partial\theta}\log f_\theta(x) \\ \frac{\partial}{\partial\theta}\log f_\theta(x) &= \frac{I(\theta)}{g'(\theta)}(T(x) - g(\theta)). \end{aligned}$$

Hence by solving the differential equation,

$$f_\theta(x) = u(\theta)\exp^{q(\theta)t(x)} v(x).$$

So, $f_\theta(x)$ is of the exponential form.

Result 2 If T is a biased estimator of $g(\theta)$ with bias $B(\theta)$ such that $B'(\theta)$ exists. Then

$$MSE_\theta(T) \geq \frac{(B'(\theta) + g'(\theta))^2}{I(\theta)} + B^2(\theta)$$

Proof: Since $E_\theta(T) = B(\theta) + g(\theta)$ it follows that

$$Cov_\theta(S,T) = \frac{\partial}{\partial\theta} E_\theta(T(\mathbf{x})) = B'(\theta) + g'(\theta).$$

In Module 2 , We have already seen that

$$MSE_\theta(T) = Var_\theta(T) + B^2(\theta)$$

By CRLB,

$$Var_\theta(T) \geq \frac{(B'(\theta) + g'(\theta))^2}{I(\theta)}.$$

Hence we get

$$MSE_\theta(T) \geq \frac{(B'(\theta) + g'(\theta))^2}{I(\theta)} + B^2(\theta).$$

Some examples:

Example 1 Let $X_1, X_2, \ldots, X_n$ be a random sample from, $f_\theta(x) = \theta \exp(-\theta x)$ for $x > 0$. Take $g(\theta) = \theta$. We are interested in finding that UE of $\frac{1}{\theta}$ whose variance attains the CRLB. Note that, $\frac{\partial}{\partial\theta} \log f_\theta(\mathbf{x}) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = n(\frac{1}{\theta} - \bar{x})$ and from (2), $\bar{X}$ is the $\frac{1}{\theta}$ whose variance attains the CRLB i.e. $\bar{X}$ is the UMVUE of $\frac{1}{\theta}$.

Example 2 Let $X_1, X_2, \ldots, X_n$ be a random sample from, $f_\lambda(x) = \exp(-\lambda)\lambda^x/x!$ for $x = 1, 2, \ldots$. Let us find the lower bound for the variance of an unbiased estimator of λ.

$$E_\lambda\left[\left(\frac{\partial}{\partial\lambda} \log f_\lambda(X)\right)^2\right] = E_\lambda\left[\left(\frac{X}{\lambda} - 1\right)^2\right] = \frac{1}{\lambda^2} var(X) = \frac{1}{\lambda}.$$

Therefore the CRLB for the variance of an unbiased estimator of λ is equal to $\frac{\lambda}{n}$.

Example 3 Let $X_1, X_2, \ldots, X_n$ be a random sample from, $N(\mu, \sigma^2)$ with μ unknown and σ known. Let us find the lower bound for the variance of an unbiased estimators of μ. Now here,

$$f_\mu(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

So,

$$\frac{\partial}{\partial \mu} \log f_\mu(x) = -\frac{2(x-\mu)}{2\sigma^2}(-1) = \frac{(x-\mu)}{\sigma^2}.$$

Hence,

$$E_\mu\left[\left(\frac{\partial}{\partial \mu} \log f_\mu(X)\right)^2\right] = E_\mu\left[\left(\frac{(x-\mu)}{\sigma^2}\right)^2\right] = \frac{var_\mu(X)}{\sigma^2} = 1/\sigma^2.$$

Therefore the lower bound for the variance of an unbiased estimators of μ is $\frac{\sigma^2}{n}$.

Some remarks on CRLB

1. If the regularity conditions are violated then the the variance of the UMVUE may be less than the CRLB.

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from $R(0,\theta)$. We know that $X_{(n)}$, the largest order statistic, is a complete sufficient statistic for θ. The density for $X_{(n)}$ is given as,

$$\begin{aligned} f_\theta(x_{(n)}) &= n\frac{x_{(n)}^{n-1}}{\theta^n}, \text{if } 0 < x_{(n)} < \theta \\ &= 0, \text{otherwise.} \end{aligned}$$

By Lehmann-Scheffe theorem a function of $X_{(n)}$ which is an unbiased estimator of θ is the UMVUE of θ. Now, $E_\theta(X_{(n)}) = \frac{n}{n+1}\theta$ and $E_\theta(X_{(n)}^2) = \frac{n}{n+2}\theta^2$. So,

$$Var_\theta(X_{(n)}) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2\theta^2 = n\frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

By Lehmann-Scheffe theorem, $T = \frac{n+1}{n}X_{(n)}$ is the UMVUE of θ. Therefore $Var(T) = \frac{\theta^2}{n(n+2)}$. The joint p.d.f. of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is given by

$$f_\theta(\mathbf{x}) = \frac{1}{\theta^n} \text{ if } \ 0 < x_i < \theta, i = 1, 2, \ldots, n$$

Since $E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right)^2\right] = \frac{n}{\theta^2}$, the CRLB of the variance of an UE of θ is equal to $\frac{\theta^2}{n}$ which is greater than the variance of the UMVUE of θ.This is so because here the the support of the density depend on the parameter θ.

2. The variance of the UMVUE may not attain CRLB.

Example $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta, 1)$. Suppose $g(\theta) = \theta^2$. The CRLB for the variance of an unbiased estimator of θ^2 is equal to $\frac{4\theta^2}{n}$. By Lehmann-Scheffe theorem, $T = \bar{X}^2 - \frac{1}{n}$ is the UMVUE of θ^2 and $Var_\theta(T) = \frac{4\theta^2}{n} + \frac{2}{n^2}$. So variance of the UMVUE may not attain CRLB.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Module- 17

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Bhattacharyya System of lower bound
In this module we shall discuss a generalization of Cramer-Rao lower bound which is known as Bhattacharya system of lower bounds. Let us consider a family of distributions $\mathcal{F}_\theta = \{f_\theta(\mathbf{x}); \theta \in \mathbf{\Theta}\}$ which satisfies the following Bhattacharya regularity conditions,

1. Θ is an open interval of the real line.
2. The support $\mathcal{X} = \{\mathbf{x} : f_\theta(\mathbf{x}) > 0\}$ does not depend on θ.
3. For some integer K, $\frac{\partial^i}{\partial\theta^i} f_\theta(\mathbf{x})$ exists for all x and for all θ for all $i = 1, \ldots, K$.
4. For any statistic $h(\mathbf{X})$ with $E_\theta(|h(\mathbf{X})|) < \infty$ for all θ, $\frac{\partial^i}{\partial\theta^i} E_\theta(h(\mathbf{X})) = \frac{\partial^i}{\partial^i\theta} \int h(\mathbf{x}) f_\theta(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) \frac{\partial^i}{\partial^i\theta} f_\theta(\mathbf{x}) d\mathbf{x}$ for all θ and for all $i = 1, \ldots, K$.
5. The matrix $\mathbf{V_K} = ((v_{ij}))$ exists for all θ and is positive definite where, $v_{ij}(\theta) = E_\theta\left[\left(\frac{\partial^i}{\partial\theta^i} \log f_\theta(\mathbf{x})\right)\left(\frac{\partial^j}{\partial\theta^j} \log f_\theta(\mathbf{x})\right)\right]$ for all $\theta \in \Theta$.

Theorem. Let the family of distribution $P_\theta = \{f_\theta(x), \theta \in \Theta\}$ satisfies the Bhattacharya regularity conditions and $g(\theta)$ be a real valued function of θ such that $g(\theta)$ is K-times differentiable for some integer K. Let T be an unbiased estimator of $g(\theta)$ such that $Var_\theta(T)$ finite for all θ , then

$$Var_\theta(T) \geq \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{g}, \text{ for all } \theta$$

where $\mathbf{g} = (g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(K)}(\theta))'$ and $g^{(i)}(\theta) = \frac{\partial^i}{\partial\theta^i} g(\theta)$.
Note :- For K=1, the Bhattacharya lower bound (and the regularity conditions) reduces to Cramer-Rao lower bound (and the corresponding regularity conditions).
Proof: Define, $S_i(\theta, \mathbf{x}) = \frac{1}{f_\theta(\mathbf{x})} \frac{\partial^i}{\partial\theta^i} f_\theta(\mathbf{x})$. Hence,

$$E(S_i) = \int \frac{1}{f_\theta(\mathbf{x})} \frac{\partial^i}{\partial\theta^i} f_\theta(\mathbf{x}) f_\theta(\mathbf{x}) d\mathbf{x}$$

$$
\begin{aligned}
&= \int \frac{\partial^i}{\partial\theta^i} f_\theta(\mathbf{x}) d\mathbf{x} \\
&= \frac{\partial^i}{\partial\theta^i} \int f_\theta(\mathbf{x}) d\mathbf{x} \\
&= \frac{\partial^i}{\partial\theta^i} 1 = 0, \forall \;\; i
\end{aligned}
$$

Therefore, $V(S_i) = v_{ii}$, $cov(S_i, S_j) = v_{ij}$, and

$$
\begin{aligned}
cov(S_i, T) &= E(S_i T) \\
&= \int t(\mathbf{x}) \frac{1}{f_\theta(\mathbf{x})} \frac{\partial^i}{\partial\theta^i} f_\theta(\mathbf{x}) f_\theta(x) d\mathbf{x} \\
&= \frac{\partial^i}{\partial\theta^i} \int t(\mathbf{x}) f_\theta(\mathbf{x}) d\mathbf{x} \\
&= \frac{\partial^i}{\partial\theta^i} g(\theta) = g^{(i)}(\theta).
\end{aligned}
$$

Define,

$$
\mathbf{\Sigma}^{K+1,K+1} = \text{ dispersion matrix of } (T, S_1, S_2, \ldots S_k) = \left(\begin{array}{c|c} var_\theta(T) & g^{(1)}(\theta) \cdots g^{(K)}(\theta) \\ \hline g^{(1)}(\theta) & \\ & \mathbf{V}_K \\ \vdots & \\ g^{(K)}(\theta) & \end{array}\right).
$$

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Since $\mathbf{\Sigma}$ is a positive definite matrix. Therefore $det(\mathbf{\Sigma}) = |\mathbf{V}||var_\theta(T) - \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{g}| \geq 0$. Hence, $var_\theta(T) \geq \mathbf{g}'\mathbf{V}_K^{-1}\mathbf{g}$.

Case of equality. Equality sign holds if $|\Sigma| = 0$ i.e. $Rank(\mathbf{\Sigma}) < K + 1$ and hence $Rank(\mathbf{\Sigma}) = K$ since $\mathbf{\Sigma}$ contains V which is non-singular. Hence there exists a non-zero vector $\mathbf{l}$ such that

$$
T - g(\theta) = \mathbf{l}'\mathbf{S}, \quad \text{with probability one where } \mathbf{S} = (S_1, S_2, \ldots S_k)'
$$

$$
\Rightarrow T - g(\theta) = \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}, \quad \text{with probability one.}
$$

Proof: Note that

$$
\begin{aligned}
Var_\theta(\mathbf{l}'\mathbf{S} - \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}) &= Var_\theta(T - g(\theta) - \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}) \\
&= Var_\theta(T) + \mathbf{g}'\mathbf{V_K}^{-1} Disp(\mathbf{S}) \mathbf{V_K}^{-1}\mathbf{g} - 2Cov(T, \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}) \\
&= \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{g} + \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{g} - 2\mathbf{g}'\mathbf{V_K}^{-1}\mathbf{g} \\
&= 0
\end{aligned}
$$

Hence, $\mathbf{l}'\mathbf{S} - \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S} = E(\mathbf{l}'\mathbf{S} - \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}) = 0,$ with probability one.
Thus we get

$$T - g(\theta) = \mathbf{l}'\mathbf{S} = \mathbf{g}'\mathbf{V_K}^{-1}\mathbf{S}, \quad \text{with probability one.}$$

Now we consider a very important result relating to Bhattacharya system of lower bounds.

Result: Suppose the n^{th} lower bound be denoted by $\Delta_n = \mathbf{g}_n'\mathbf{V}_n^{-1}\mathbf{g}_n$, where $\mathbf{g}_n' = (g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(n)}(\theta))'$.The sequence of lower bounds $\{\Delta_n\}$ is non-decreasing sequence, *i.e.*, $\Delta_{n+1} \geq \Delta_n$. Observe that, Δ_1 is Cramer-Rao lower bound.

Proof Note that, $\Delta_{n+1} = \mathbf{g}_{n+1}'\mathbf{V}_{n+1}^{-1}\mathbf{g}_{n+1}$, and

$$\mathbf{g}_{n+1}' = (g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(n)}(\theta) \vdots g^{(n+1)}(\theta))'.$$

Suppose the vector $\mathbf{g}_{n+1}'$ and the matrix $\mathbf{V}_{n+1}$ be partitioned as follows $\mathbf{g}_{n+1}' = (\mathbf{g}_n' \vdots g^{(n+1)}(\theta))$,

$$\mathbf{V}_{n+1} = \left(\begin{array}{c|c} \mathbf{V}_n & \mathbf{u}_{n+1} \\ \hline \mathbf{u}_{n+1}' & v_{n+1,n+1} \end{array}\right).$$

For any non-singular matrix C we can write,

$$\begin{aligned} \Delta_{n+1} &= \mathbf{g}_{n+1}'C'C'^{-1}\mathbf{V}_{n+1}^{-1}C'^{-1}C'\mathbf{g}_{n+1} \\ &= (C\mathbf{g}_{n+1}')\left(C\mathbf{V}_{n+1}C'^{-1}\right)(C\mathbf{g}_{n+1}). \end{aligned}$$

If we choose,

$$C = \left(\begin{array}{c|c} \mathbf{I}_n & 0 \\ \hline -\mathbf{u}_{n+1}'\mathbf{V}_n & 1 \end{array}\right),$$

then we have $C\mathbf{g}_{n+1} = (\mathbf{g}_n, g^{(n+1)} - \mathbf{u}_{n+1}'\mathbf{V}_n\mathbf{g}_n)'$, and

$$\begin{aligned} C\mathbf{V}_{n+1}C' &= \left(\begin{array}{c|c} \mathbf{I}_n & 0 \\ \hline -\mathbf{u}_{n+1}'\mathbf{V}_n & 1 \end{array}\right)\left(\begin{array}{c|c} \mathbf{V}_n & \mathbf{u}_{n+1} \\ \hline \mathbf{u}_{n+1}' & v_{n+1,n+1} \end{array}\right)\left(\begin{array}{c|c} \mathbf{I}_n & -\mathbf{u}_{n+1}'\mathbf{V}_n \\ \hline 0 & 1 \end{array}\right) \\ &= \left(\begin{array}{c|c} \mathbf{V}_n & \mathbf{v}_{n+1} \\ \hline \mathbf{u}_{n+1}' - \mathbf{u}_{n+1}' & v_{n+1,n+1} - \mathbf{u}_{n+1}'\mathbf{V}_n\mathbf{u}_{n+1} \end{array}\right)\left(\begin{array}{c|c} \mathbf{I}_n & -\mathbf{u}_{n+1}'\mathbf{V}_n \\ \hline 0 & 1 \end{array}\right) \\ &= \left(\begin{array}{c|c} \mathbf{V}_n & \mathbf{v}_{n+1}' - \mathbf{u}_{n+1}' \\ \hline \mathbf{u}_{n+1}' - \mathbf{u}_{n+1}' & v_{n+1,n+1} - \mathbf{u}_{n+1}'\mathbf{V}_n\mathbf{u}_{n+1} \end{array}\right) \\ &= \left(\begin{array}{c|c} \mathbf{V}_n & 0 \\ \hline 0 & E_{n+1,n+1} \end{array}\right) \end{aligned}$$

where $E_{n+1,n+1} = v_{n+1,n+1} - \mathbf{u}'_{n+1}\mathbf{V}_n\mathbf{u}_{n+1}$. Since, $\mathbf{V}_{n+1}$ is positive definite matrix, $\mathbf{V}_n$ is p.d. and $E_{n+1,n+1} > 0$. Therefore,

$$(C\mathbf{V}_{n+1}C')^{-1} = \left(\begin{array}{c|c} \mathbf{V}_n^{-1} & 0 \\ \hline 0 & \frac{1}{E_{n+1,n+1}} \end{array}\right).$$

So finally,

$$\begin{aligned}
\Delta_{n+1} &= (C\mathbf{g}_{n+1}{}') \left(C\mathbf{V}_{n+1}(\theta)C'^{-1}\right)(C\mathbf{g}_{n+1}) \\
&= \mathbf{g}'_n\mathbf{V}_n^{-1}(\theta)\mathbf{g}_n + \frac{\left(g^{(n+1)} - \mathbf{u}'_{n+1}\mathbf{V}_n\mathbf{g}_n\right)^2}{E_{n+1,n+1}} \\
&\geq \Delta_n.
\end{aligned}$$

Note : The implication of the result that if there is no UE of $g(\theta)$ which attains the n th lower bound Δ_n, then a sharper lower bound Δ_{n+1} can be obtained and looked at. If an UE satisfies the n th lower bound Δ_n then no further improvement can be made and hence $\Delta_n = \Delta_{n+1}$. However $\Delta_n = \Delta_{n+1}$ does not imply that there exists an UE of $g(\theta)$ which attains the n th lower bound. Consider the following example:

Example : Let $X_1, X_2, \ldots, X_n$ be independent sample from $N(\theta, 1)$. In the last module we have seen that there does not exists an UE of θ^2 which attains Δ_1 i.e. the CRLB. We want to find an UE of θ^2 which attains the Bhattacharya lower bound Δ_2.
The joint p.d.f. of $\mathbf{X}$ is given by,

$$f_\theta(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n}\exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right\}.$$

Hence,

$$\begin{aligned}
S_1 &= \frac{1}{f_\theta(\mathbf{x})}\frac{\partial}{\partial\theta}f_\theta(\mathbf{x}) \\
&= \frac{1}{f_\theta(\mathbf{x})}f_\theta(\mathbf{x})\sum_{i=1}^n (x_i - \theta) \\
&= \sum_{i=1}^n (x_i - \theta) \\
&= n(\bar{x} - \theta)
\end{aligned}$$

A Gateway to All Post Graduate Courses

and

$$
\begin{aligned}
S_2 &= \frac{1}{f_\theta(\mathbf{x})}\frac{\partial^2}{\partial\theta^2}f_\theta(\mathbf{x}) \\
&= \frac{1}{f_\theta(x)}\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}f_\theta(x)\right) \\
&= \frac{1}{f_\theta(x)}\frac{\partial}{\partial\theta}\left(f_\theta(x)\sum_{i=1}^{n}(x_i-\theta)\right) \\
&= \frac{1}{f_\theta(x)}\left(\frac{\partial}{\partial\theta}f_\theta(x)\sum_{i=1}^{n}(x_i-\theta)+f_\theta(x)(-n)\right) \\
&= \frac{1}{f_\theta(x)}\left(f_\theta(x)\left(\sum_{i=1}^{n}(x_i-\theta)\right)^2+f_\theta(x)(-n)\right) \\
&= \left(\sum_{i=1}^{n}(x_i-\theta)\right)^2-n \\
&= n^2(\bar{x}-\theta)^2-n.
\end{aligned}
$$

By definition we know, $E(S_1)=E(S_2)=0$. Therefore, $E(S_1^2)=E\left[\{n(\bar{x}-\theta)\}^2\right]=n^2\frac{1}{n}=n$,

$$
\begin{aligned}
E(S_2^2) &= E\left[n^4\left(\bar{X}-\theta)\right)^2+n^2-2n^3(\bar{X}-\theta)^2\right] \\
&= n^4 3\frac{1}{n^2}+n^2-2n^3\frac{1}{n} \\
&= 2n^2,
\end{aligned}
$$

and $E(S_1S_2)=n^2E[(\bar{X}-\theta)^3]-n^2E(\bar{X}-\theta)=0$. Therefore,

$$
\mathbf{V}_2=\left(\begin{array}{c|c} n & 0 \\ \hline 0 & 2n^2 \end{array}\right).
$$

Since the parameter of interest under consideration is $g(\theta)=\theta^2$ we have, $\mathbf{g}'=(2\theta,2)$. Therefore, the V_2, BLB for θ^2 is given by,

$$
\begin{aligned}
\Delta_2 &= \mathbf{g}'V^{-1}\mathbf{g} \\
&= \frac{4}{n}(\theta,1)\left(\begin{array}{c|c} \frac{1}{n} & 0 \\ \hline 0 & \frac{1}{2n^2} \end{array}\right)\left(\begin{array}{c} \theta \\ 1 \end{array}\right)
\end{aligned}
$$

e Pathshala पाठशाला A Gateway to All Post Graduate Courses

$$
\begin{aligned}
&= \frac{4}{n}\left(\theta, \frac{1}{2n}\right)\begin{pmatrix}\theta \\ 1\end{pmatrix} \\
&= \frac{4}{n}\left(\theta^2 + \frac{1}{2n}\right)
\end{aligned}
$$

The UE of θ^2 which attains the Bhattacharya lower bound of order 2 is given by

$$
\begin{aligned}
\mathbf{g}'\mathbf{V_2}^{-1}\mathbf{S} &= \mathbf{g}'V^{-1}\mathbf{g} \\
&= (2\theta, 2)\left(\begin{array}{c|c} \frac{1}{n} & 0 \\ \hline 0 & \frac{1}{2n^2}\end{array}\right)\begin{pmatrix} n(\bar{x}-\theta) \\ n^2(\bar{x}-\theta)^2 - n\end{pmatrix} \\
&= \left(\frac{2\theta}{n}, \frac{1}{n^2}\right)\begin{pmatrix} n(\bar{x}-\theta) \\ n^2(\bar{x}-\theta)^2 - n\end{pmatrix} \\
&= \left(\bar{x}^2 - \frac{1}{n}\right) - \theta^2
\end{aligned}
$$

Hence $T = \bar{X}^2 - \frac{1}{n}$ is the unbiased estimator of θ^2 which attains Bhattacharya lower bound of order 2.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Chapman-Robbins Lower Bound

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In Module-16 we have discussed the Cramer-Rao lower bound to the variance of an unbiased estimator os some parametric function. But the problem in applying the Cramer-Rao lower bound is that in all the cases the underlying assumptions does not hold true and therefore we can not apply this bound in all situations. To overcome this problem Chapman and Robbins (1951) have provided a lower bound to the variance of an unbiased estimator. The importance of this inequality is that it is derived without making various assumptions that was done in the case of Cramer-Rao Inequality. We shall suppose that the parameter space Θ is an open sub interval of the real line.

Chapman-Robbins Inequality Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from $f_\theta(\cdot)$, where θ belongs to Θ. Let $T = t(\mathbf{X})$ be an unbiased estimator of $g(\theta)$. We will consider the case where $f_\theta(\cdot)$ is a probability density function; the development for discrete density functions is analogous.Let $\theta_0 \in \Theta$ be any fixed value of θ, such that for sufficiently small $h \neq 0$,$\theta_0 + h \in \Theta$ and $f_{\theta_0}(\mathbf{x}) = 0 \Rightarrow f_{\theta_0+h}(\mathbf{x}) = 0$. Then

$$Var_{\theta_0}(T) \geq Sup_h \left[\frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} - 1\}^2} \right]. \tag{1}$$

Note: The quantity in the right-hand side of the above inequality is called the Chapman-Robbins lower bound for the variance of an unbiased estimator of $g(\theta)$.

Proof Since T is an unbiased estimator of $g(\theta)$ we have

$$E_\theta(T) = g(\theta), \forall \theta$$

$$\Rightarrow \int t f_\theta(\mathbf{x}) d\mathbf{x} = g(\theta), \forall \theta. \tag{2}$$

From (2) and using the fact that $\int\{f_{\theta_0+h}(\mathbf{x})d\mathbf{x} - f_{\theta_0}(\mathbf{x})\}d\mathbf{x} = 0$, we get

$$\begin{aligned} g(\theta_0+h) - g(\theta_0) &= \int\{tf_{\theta_0+h}(\mathbf{x}) - tf_{\theta_0}(\mathbf{x})\}d\mathbf{x} \\ &= \int\{t - g(\theta_0)\}\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}d\mathbf{x} \\ &= \int\{t - g(\theta_0)\}\frac{\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}}{f_{\theta_0}(\mathbf{x})}f_{\theta_0}(\mathbf{x})d\mathbf{x} \end{aligned}$$

This means

$$Cov\left[\{T - g(\theta_0)\}, \frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right] = g(\theta_0+h) - g(\theta_0).$$

So, using the well known result that for any two random variables U and V

$$\{Cov(U,V)\}^2 \leq Var(U)Var(V),$$

we see that

$$\{g(\theta_0+h) - g(\theta_0)\}^2 \leq Var_{\theta_0}(T)Var_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right]. \qquad (3)$$

Since $E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X})-f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right] = \int\frac{\{f_{\theta_0+h}(\mathbf{x})-f_{\theta_0}(\mathbf{x})\}}{f_{\theta_0}(\mathbf{x})}f_{\theta_0}(\mathbf{x})d\mathbf{x} = 0$, from (3) we get

$$Var_{\theta_0}(T) \geq \frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X})-f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right]^2} = \frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\right]^2}. \qquad (4)$$

Since (4) holds for all values of h, we get

$$Var_{\theta_0}(T) \geq Sup_h\left[\frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\{\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\}^2}\right].$$

Remark 1 The denominator of quantity in the R.H.S. of (1)can be written as

$$\begin{aligned} E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\right]^2 &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 2E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right] + 1 \\ &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 2\int f_{\theta_0+h}(\mathbf{x})d\mathbf{x} + 1 \\ &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 1 \end{aligned}$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

So the inequality can be rewritten as

$$Var_{\theta_0}(T) \geq Sup_h \left[\frac{\{g(\theta_0+h)-g(\theta_0)\}^2}{E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 - 1} \right].$$

Remark 2 If we assume that $g(\theta)$ possesses a derivative at θ_0, then dividing the numerator and denominator R.H.S. of of (1) by h^2 and taking limit as $h \to 0$, we get

$$Var_{\theta_0}(T) \geq \frac{\{g'(\theta_0)\}^2}{lim_{h\to 0} E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X})-f_{\theta_0}(\mathbf{X})\}}{hf_{\theta_0}(\mathbf{X})}\right]^2}. \tag{5}$$

If, in addition, $f_\theta(\mathbf{X})$ is differentiable w.r.t.θ and the derivative is continuous in a certain neighbourhood of θ_0, then the denominator of the R.H.S,of (5) becomes

$$\begin{aligned}
lim_{h\to 0} E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X})-f_{\theta_0}(\mathbf{X})\}}{hf_{\theta_0}(\mathbf{X})}\right]^2 &= lim_{h\to 0}\int\left[\frac{\{f_{\theta_0+h}(\mathbf{x})-f_{\theta_0}(\mathbf{x})\}}{hf_{\theta_0}(\mathbf{x})}\right]^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int lim_{h\to 0}\left[\frac{\{f_{\theta_0+h}(\mathbf{x})-f_{\theta_0}(\mathbf{x})\}}{hf_{\theta_0}(\mathbf{x})}\right]^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int\left(\frac{[\frac{\partial f_\theta(\mathbf{x})}{\partial\theta}]_{\theta_0}}{f_{\theta_0}(\mathbf{x})}\right)^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int\left(\left[\frac{\partial ln f_\theta(\mathbf{x})}{\partial\theta}\right]_{\theta_0}\right)^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= E_{\theta_0}\left(\left[\frac{\partial ln f_\theta(\mathbf{x})}{\partial\theta}\right]_{\theta_0}\right)^2
\end{aligned}$$

and the Chapman-Robbins inequality becomes identical with the Cramer-Rao inequality. The point to be noted here that even two extra assumptions needed in this case imposes no restriction on the nature of T. It may be regarded as an improvement on the Cramer-Rao inequality because it does not require the stringent regularity conditions that underlie the latter.

e-pathshala पाठशाला A Gateway to All Post Graduate Courses

Example Let $X_1, X_2, \ldots X_n$ be a random sample of size n from a rectangular distribution with p.d.f.

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta} \text{ if } 0 < x < \theta \\ &= 0 \text{ otherwise,} \end{aligned}$$

where $0 < \theta < \infty$. In module-16 we have seen that as the support of the distribution depends on θ, the variance of the UMVUE of θ falls below the the Cramer-Rao lower bound. So in this situation one may wish to find the Chapman-Robbins lower bound to variance of an unbiased estimator of θ.The joind p.d.f. of $\mathbf{X} = (X_1, X_2, \ldots X_n)$ is given by

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta^n} \text{ if } 0 < x_i < \theta \text{ for all } i \\ &= 0 \text{ otherwise,} \end{aligned}$$

In this context the condition $f_{\theta_0}(\mathbf{x}) = 0 \Rightarrow f_{\theta_0+h}(\mathbf{x}) = 0$ implies that if an observation $X_i > \theta_0$ then $X_i > \theta_0 + h$ which is possible only when $h < 0$. But the condition $\theta_0 + h \in \Theta$ implies that $\theta_0 + h > 0$. So $-\theta_0 < h < 0$. If we take $g(\theta) = \theta$, then by the Chapman-Robbins lower bound to variance of an unbiased estimator of θ is

$$Sup_h \left[\frac{h^2}{E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 - 1} \right]$$

Now

$$\begin{aligned} \frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} &= \frac{\theta_0^n}{(\theta_0+h)^n} \text{ if } 0 < x_i < \theta_0 + h \text{ for all } i \\ &= 0 \text{ otherwise.} \end{aligned}$$

So

$$E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 = \int \left[\frac{\theta_0^n}{(\theta_0+h)^n}\right]^2 \frac{1}{\theta_0^n} d\mathbf{x},$$

where the n-fold integral is taken over the intervals $0 < x_i < \theta_0 + h$ for all i. Hence we get

$$E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 = \left[\frac{\theta_0^n}{(\theta_0+h)^n}\right]^2 \frac{1}{\theta_0^n}(\theta_0+h)^n = \frac{\theta_0^n}{(\theta_0+h)^n}.$$

Hence the the Chapman-Robbins lower bound to the variance of an unbiased estimator of θ is

$$Sup_{-\theta_0<h<0}\frac{h^2(\theta_0+h)^n}{\theta_0^n-(\theta_0+h)^n}.$$

In particular when $n = 1$, it reduces to

$$Sup_{-\theta_0<h<0}\{-h(\theta_0+h)\}.$$

To calculate the supremum note that

$$-h(\theta_0+h) = \frac{\theta_0^2}{4} - \left(h+\frac{\theta_0}{2}\right)^2 \leq \frac{\theta_0^2}{4}.$$

Hence $Sup_{-\theta_0<h<0}\{-h(\theta_0+h)\} = \frac{\theta_0^2}{4}$ and the supremum value is attained when $h = \frac{-\theta_0}{2}$. Since θ_0 is arbitrary, the Chapman-Robbins lower bound to the variance of an unbiased estimator of θ is $\frac{\theta^2}{4}$ if $n = 1$.

Remark By Lehmann-Scheffe theorem, $T = \frac{n+1}{n}X_{(n)}$ is the UMVUE of θ and in Module 16 we have seen that with $Var_\theta(T) = \frac{\theta^2}{n(n+2)}$ which becomes $\frac{\theta^2}{3}$ when $n = 1$. Hence $Var_\theta(T)$ exceeds the Chapman-Robbins lower bound $\frac{\theta^2}{4}$ but it is less than the Cramer-Rao lower bound θ^2 that we have already seen in Module-16. This apparent discrepancy occurs due to the fact that in this case the support of the distribution depends on θ and therefore the Cramer-Rao lower bound is inappropriate for this problem. So the Chapman-Robbins lower bound is an appropriate lower bound for the this problem.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Chapman-Robbins Lower Bound

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In Module-16 we have discussed the Cramer-Rao lower bound to the variance of an unbiased estimator os some parametric function. But the problem in applying the Cramer-Rao lower bound is that in all the cases the underlying assumptions does not hold true and therefore we can not apply this bound in all situations. To overcome this problem Chapman and Robbins (1951) have provided a lower bound to the variance of an unbiased estimator. The importance of this inequality is that it is derived without making various assumptions that was done in the case of Cramer-Rao Inequality. We shall suppose that the parameter space Θ is an open sub interval of the real line.

Chapman-Robbins Inequality Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from $f_\theta(\cdot)$, where θ belongs to Θ. Let $T = t(\mathbf{X})$ be an unbiased estimator of $g(\theta)$. We will consider the case where $f_\theta(\cdot)$ is a probability density function; the development for discrete density functions is analogous.Let $\theta_0 \in \Theta$ be any fixed value of θ, such that for sufficiently small $h \neq 0$,$\theta_0 + h \in \Theta$ and $f_{\theta_0}(\mathbf{x}) = 0 \Rightarrow f_{\theta_0+h}(\mathbf{x}) = 0$. Then

$$Var_{\theta_0}(T) \geq Sup_h \left[\frac{\{g(\theta_0 + h) - g(\theta_0)\}^2}{E_{\theta_0}\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} - 1\}^2} \right]. \tag{1}$$

Note: The quantity in the right-hand side of the above inequality is called the Chapman-Robbins lower bound for the variance of an unbiased estimator of $g(\theta)$.

Proof Since T is an unbiased estimator of $g(\theta)$ we have

$$E_\theta(T) = g(\theta), \forall\theta$$

$$\Rightarrow \int t f_\theta(\mathbf{x}) d\mathbf{x} = g(\theta), \forall\theta. \tag{2}$$

From (2) and using the fact that $\int\{f_{\theta_0+h}(\mathbf{x})d\mathbf{x} - f_{\theta_0}(\mathbf{x})\}d\mathbf{x} = 0$, we get

$$\begin{aligned} g(\theta_0+h) - g(\theta_0) &= \int\{tf_{\theta_0+h}(\mathbf{x}) - tf_{\theta_0}(\mathbf{x})\}d\mathbf{x} \\ &= \int\{t - g(\theta_0)\}\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}d\mathbf{x} \\ &= \int\{t - g(\theta_0)\}\frac{\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}}{f_{\theta_0}(\mathbf{x})}f_{\theta_0}(\mathbf{x})d\mathbf{x} \end{aligned}$$

This means

$$Cov\left[\{T - g(\theta_0)\}, \frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right] = g(\theta_0+h) - g(\theta_0).$$

So, using the well known result that for any two random variables U and V

$$\{Cov(U,V)\}^2 \le Var(U)Var(V),$$

we see that

$$\{g(\theta_0+h) - g(\theta_0)\}^2 \le Var_{\theta_0}(T)Var_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right]. \tag{3}$$

Since $E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right] = \int \frac{\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}}{f_{\theta_0}(\mathbf{x})}f_{\theta_0}(\mathbf{x})d\mathbf{x} = 0$, from (3) we get

$$Var_{\theta_0}(T) \ge \frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{f_{\theta_0}(\mathbf{X})}\right]^2} = \frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\right]^2}. \tag{4}$$

Since (4) holds for all values of h, we get

$$Var_{\theta_0}(T) \ge Sup_h\left[\frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\{\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\}^2}\right].$$

Remark 1 The denominator of quantity in the R.H.S. of (1)can be written as

$$\begin{aligned} E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} - 1\right]^2 &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 2E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right] + 1 \\ &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 2\int f_{\theta_0+h}(\mathbf{x})d\mathbf{x} + 1 \\ &= E_{\theta_0}\left[\frac{f_{\theta_0+h}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}\right]^2 - 1 \end{aligned}$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

So the inequality can be rewritten as

$$Var_{\theta_0}(T) \geq Sup_h \left[\frac{\{g(\theta_0+h) - g(\theta_0)\}^2}{E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 - 1} \right].$$

Remark 2 If we assume that $g(\theta)$ possesses a derivative at θ_0, then dividing the numerator and denominator R.H.S. of of (1) by h^2 and taking limit as $h \to 0$, we get

$$Var_{\theta_0}(T) \geq \frac{\{g'(\theta_0)\}^2}{lim_{h\to 0} E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{hf_{\theta_0}(\mathbf{X})}\right]^2}. \tag{5}$$

If, in addition, $f_\theta(\mathbf{X})$ is differentiable w.r.t.θ and the derivative is continuous in a certain neighbourhood of θ_0, then the denominator of the R.H.S,of (5) becomes

$$\begin{aligned}
lim_{h\to 0} E_{\theta_0}\left[\frac{\{f_{\theta_0+h}(\mathbf{X}) - f_{\theta_0}(\mathbf{X})\}}{hf_{\theta_0}(\mathbf{X})}\right]^2 &= lim_{h\to 0} \int \left[\frac{\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}}{hf_{\theta_0}(\mathbf{x})}\right]^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int lim_{h\to 0} \left[\frac{\{f_{\theta_0+h}(\mathbf{x}) - f_{\theta_0}(\mathbf{x})\}}{hf_{\theta_0}(\mathbf{x})}\right]^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int \left(\frac{[\frac{\partial f_\theta(\mathbf{x})}{\partial\theta}]_{\theta_0}}{f_{\theta_0}(\mathbf{x})}\right)^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= \int \left(\left[\frac{\partial ln f_\theta(\mathbf{x})}{\partial\theta}\right]_{\theta_0}\right)^2 f_{\theta_0}(\mathbf{x})d\mathbf{x} \\
&= E_{\theta_0}\left(\left[\frac{\partial ln f_\theta(\mathbf{x})}{\partial\theta}\right]_{\theta_0}\right)^2
\end{aligned}$$

and the Chapman-Robbins inequality becomes identical with the Cramer-Rao inequality. The point to be noted here that even two extra assumptions needed in this case imposes no restriction on the nature of T. It may be regarded as an improvement on the Cramer-Rao inequality because it does not require the stringent regularity conditions that underlie the latter.

e-pathshala पाठशाला A Gateway to All Post Graduate Courses

Example Let $X_1, X_2, \dots X_n$ be a random sample of size n from a rectangular distribution with p.d.f.

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta} \text{ if } 0 < x < \theta \\ &= 0 \text{ otherwise,} \end{aligned}$$

where $0 < \theta < \infty$. In module-16 we have seen that as the support of the distribution depends on θ, the variance of the UMVUE of θ falls below the the Cramer-Rao lower bound. So in this situation one may wish to find the Chapman-Robbins lower bound to variance of an unbiased estimator of θ.The joind p.d.f. of $\mathbf{X} = (X_1, X_2, \dots X_n)$ is given by

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta^n} \text{ if } 0 < x_i < \theta \text{ for all } i \\ &= 0 \text{ otherwise,} \end{aligned}$$

In this context the condition $f_{\theta_0}(\mathbf{x}) = 0 \Rightarrow f_{\theta_0+h}(\mathbf{x}) = 0$ implies that if an observation $X_i > \theta_0$ then $X_i > \theta_0 + h$ which is possible only when $h < 0$. But the condition $\theta_0 + h \in \Theta$ implies that $\theta_0 + h > 0$. So $-\theta_0 < h < 0$. If we take $g(\theta) = \theta$, then by the Chapman-Robbins lower bound to variance of an unbiased estimator of θ is

$$Sup_h \left[\frac{h^2}{E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 - 1} \right]$$

Now

$$\begin{aligned} \frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} &= \frac{\theta_0^n}{(\theta_0+h)^n} \text{ if } 0 < x_i < \theta_0 + h \text{ for all } i \\ &= 0 \text{ otherwise.} \end{aligned}$$

So

$$E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 = \int \left[\frac{\theta_0^n}{(\theta_0+h)^n}\right]^2 \frac{1}{\theta_0^n} d\mathbf{x},$$

where the n-fold integral is taken over the intervals $0 < x_i < \theta_0 + h$ for all i. Hence we get

$$E_{\theta_0}\left\{\frac{f_{\theta_0+h}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right\}^2 = \left[\frac{\theta_0^n}{(\theta_0+h)^n}\right]^2 \frac{1}{\theta_0^n}(\theta_0+h)^n = \frac{\theta_0^n}{(\theta_0+h)^n}.$$

Hence the the Chapman-Robbins lower bound to the variance of an unbiased estimator of θ is

$$Sup_{-\theta_0<h<0}\frac{h^2(\theta_0+h)^n}{\theta_0^n-(\theta_0+h)^n}.$$

In particular when $n = 1$, it reduces to

$$Sup_{-\theta_0<h<0}\{-h(\theta_0+h)\}.$$

To calculate the supremum note that

$$-h(\theta_0+h) = \frac{\theta_0^2}{4} - \left(h+\frac{\theta_0}{2}\right)^2 \leq \frac{\theta_0^2}{4}.$$

Hence $Sup_{-\theta_0<h<0}\{-h(\theta_0+h)\} = \frac{\theta_0^2}{4}$ and the supremum value is attained when $h = \frac{-\theta_0}{2}$. Since θ_0 is arbitrary, the Chapman-Robbins lower bound to the variance of an unbiased estimator of θ is $\frac{\theta^2}{4}$ if $n = 1$.

Remark By Lehmann-Scheffe theorem, $T = \frac{n+1}{n}X_{(n)}$ is the UMVUE of θ and in Module 16 we have seen that with $Var_\theta(T) = \frac{\theta^2}{n(n+2)}$ which becomes $\frac{\theta^2}{3}$ when $n = 1$. Hence $Var_\theta(T)$ exceeds the Chapman-Robbins lower bound $\frac{\theta^2}{4}$ but it is less than the Cramer-Rao lower bound θ^2 that we have already seen in Module-16. This apparent discrepancy occurs due to the fact that in this case the support of the distribution depends on θ and therefore the Cramer-Rao lower bound is inappropriate for this problem. So the Chapman-Robbins lower bound is an appropriate lower bound for the this problem.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Cramer-Rao lower bound in case of several parameters

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In this module we consider a generalization of Cramer-Rao when there are more than one parameter. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from $f_{\boldsymbol{\theta}}(\cdot)$, where the vector valued parameter $\boldsymbol{\theta}$ consists of k elements $(\theta_1, \theta_2, \ldots \theta_k)$ belonging to the parameter space Θ. Consider a vector of linearly independent parametric functions $(g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots g_k(\boldsymbol{\theta}))$ over the parameter space Θ. Let $\mathbf{T} = (T_1, T_2, \ldots T_k)$ be a vector valued statistic such that T_i is an unbiased estimator of $g_i(\boldsymbol{\theta})$ for all $i = 1, 2, \ldots k$. We assume that the variance-covariance matrix $\Sigma_{\mathbf{T}}$ of $\mathbf{T}$ exists and is positive definite. Suppose $\mathcal{F}_{\boldsymbol{\theta}} = \{f_\theta(\mathbf{x}) : \theta \in \boldsymbol{\Theta}\}$ be the family of joint p.d.f.'s of $\mathbf{X}$. It satisfies the following regularity conditions:

1. Θ is an open sub interval of the real line.

2. The support $\mathcal{X} = \{\mathbf{x} : f_\theta(\mathbf{x}) > 0\}$ does not depend on θ.

3. $\frac{\partial}{\partial\theta_i} \log f_{\boldsymbol{\theta}}(\mathbf{x})$ exists for all $\mathbf{x}$ and for all $\theta_i, i = 1, 2, \ldots k$.

4. For any statistic $h(\mathbf{X})$ with $E_{\boldsymbol{\theta}}(|h(\mathbf{X})|) < \infty$ for all θ, the the operations of integration and differentiation with respect to θ_i for all $i = 1, 2, \ldots k$ can be interchanged in $E_{\boldsymbol{\theta}}(h(\mathbf{X}))$. That is, $\frac{\partial}{\partial\theta_i} E_{\boldsymbol{\theta}}(h(\mathbf{X})) = \frac{\partial}{\partial\theta_i} \int h(\mathbf{x}) f_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) \frac{\partial}{\partial\theta_i} f_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x}$ for all $\theta_i, i = 1, 2, \ldots k$.

5. The elements of the matrix $D = (d_{ij})$, where $d_{ij} = \frac{\partial}{\partial\theta_j} g_i(\boldsymbol{\theta}), i = 1, 2, \ldots k, j = 1, 2, \ldots k$ exists and is non-singular.

6. The elements of the information matrix $\Delta = (\delta_{ij})$, where

$$\delta_{ij} = E_{\boldsymbol{\theta}} \left[\left(\frac{\partial}{\partial\theta_i} \log f_{\boldsymbol{\theta}}(\mathbf{x}) \right) \left(\frac{\partial}{\partial\theta_j} \log f_{\boldsymbol{\theta}}(\mathbf{x}) \right) \right],$$

exist and are such that Δ is positive definite.

Theorem $\mathcal{F}_\theta$ satisfies the above regularity conditions then the matrix $\Sigma_{\mathbf{T}} - D\Delta^{-1}D'$ is n.n.d.

Proof Let $\mathbf{S} = (S_1, S_2, \ldots, S_k)$ be the vector of score functions where

$$S_i = \frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\mathbf{x}), i = 1, 2, \ldots, k.$$

Then for each $i = 1, 2, \ldots k$,

$$\begin{aligned}
E_{\boldsymbol{\theta}}(S_i) &= \int \frac{1}{f_{\boldsymbol{\theta}}(\mathbf{x})}\frac{\partial}{\partial\theta_i}f_{\boldsymbol{\theta}}(\mathbf{x})f_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x} \\
&= \int \frac{\partial}{\partial\theta_i}f_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x} \\
&= \frac{\partial}{\partial\theta_i}\int f_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x} \\
&= \frac{\partial}{\partial\theta_i}1 = 0,
\end{aligned}$$

and for all $i < j = 1, 2, \ldots k$,

$$Cov_{\boldsymbol{\theta}}(S_i, S_j) = E_{\boldsymbol{\theta}}(S_iS_j) = E_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\mathbf{x})\right)\left(\frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(\mathbf{x})\right)\right] = \delta_{ij}.$$

Hence the variance-covariance matrix of $\mathbf{S}$ is equal to Δ. Also for all $i < j = 1, 2, \ldots k$,

$$\begin{aligned}
Cov_{\boldsymbol{\theta}}(T_i, S_j) &= E_{\boldsymbol{\theta}}(T_iS_j) \\
&= \int t(\mathbf{x})\frac{1}{f_{\boldsymbol{\theta}}(\mathbf{x})}\frac{\partial}{\partial\theta_j}f_{\boldsymbol{\theta}}(\mathbf{x})f_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x} \\
&= \frac{\partial}{\partial\theta_j}\int t(\mathbf{x})f_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x} \\
&= \frac{\partial}{\partial\theta_j}E_{\boldsymbol{\theta}}(T(\mathbf{x})) = \frac{\partial}{\partial\theta_j}g_i(\boldsymbol{\theta}) = d_{ij}.
\end{aligned}$$

Now for any two fixed vectors $\mathbf{u} \in \mathcal{R}^{\mathbf{k}}$ and $\mathbf{v} \in \mathcal{R}^{\mathbf{k}}$ let us define two random variables Y and Z such that

$$Y = \mathbf{u}'\mathbf{S}, \mathbf{Z} = \mathbf{v}'\mathbf{T}.$$

Then we get

$$Cov_{\boldsymbol{\theta}}(Y, Z) = E_{\boldsymbol{\theta}}(\mathbf{u}'\mathbf{S}\mathbf{T}'\mathbf{v}) = \mathbf{u}'\mathbf{D}\mathbf{v},$$

$$Var_{\boldsymbol{\theta}}(Y) = \mathbf{u}'\boldsymbol{\Delta}\mathbf{u}, \text{ and } \mathbf{Var}_{\boldsymbol{\theta}}(\mathbf{Z}) = \mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{v}.$$

Then by Cauchy-Schwarz inequality

$$(\mathbf{u}'\mathbf{D}\mathbf{v})^{\mathbf{2}} \leq (\mathbf{u}'\boldsymbol{\Delta}\mathbf{u})(\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{v}). \tag{1}$$

If we set $\mathbf{u} = \boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v}$ then as $(\Delta^{-1})' = \Delta^{-1}$, then from (1) we get

$$(\mathbf{v}'\mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v})^{\mathbf{2}} \leq (\mathbf{v}'\mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v})(\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{v}) = (\mathbf{v}'\mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v})(\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{v}). \tag{2}$$

However since Δ is pd and D is non-singular we have $(\mathbf{v}'\mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v}) \geq \mathbf{0}$ and from (2) we get

$$(\mathbf{v}'\mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}'\mathbf{v}) \leq (\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{v})$$
$$\text{or } \mathbf{v}'(\boldsymbol{\Sigma}_{\mathbf{T}} - \mathbf{D}\boldsymbol{\Delta}^{-\mathbf{1}}\mathbf{D}')\mathbf{v} \geq \mathbf{0}, \ \forall \ \mathbf{v} \in \mathcal{R}^{\mathbf{k}}.$$

Thus $\Sigma_{\mathbf{T}} - D\Delta^{-1}D'$ is n.n.d..

Remark 1 For the ith component of $\mathbf{T}$

$$Var_{\boldsymbol{\theta}}(T_i) \geq \sum_{j=1}^{k}\sum_{l=1}^{k} d_{ij}\delta^{jl}d_{il},$$

where δ^{jl} denotes the $(j,l)-th$ element of Δ^{-1}. In particular, when $k=1$ we have

$$Var_{\theta_1}(T_1) \geq (d_{11})^2\delta^{11} = \left\{\frac{\partial}{\partial\theta_1}g_1(\theta_1)\right\}^2 / E_{\theta_1}\left[\left(\frac{\partial}{\partial\theta_1}\log f_{\theta_1}(\mathbf{x})\right)^2\right].$$

The R.H.S. of the above inequality is Cramer-Rao lower bound to the variance of an unbiased estimator T_1 of $g_1(\theta_1)$.

Remark 2 If $g = (\theta_1, \theta_2, \ldots \theta_k)$, we have $D = I_k$ where I_k denotes the identity matrix of order k. Then the theorem states that $\Sigma_{\mathbf{T}} - \Delta^{-1}$ is n.n.d..

Remark 3 Since $\Sigma_{\mathbf{T}} - D\Delta^{-1}D'$ is n.n.d. and $D\Delta^{-1}D'$ is also n.n.d. it follows that

$$|\Sigma_{\mathbf{T}}| \geq |D\Delta^{-1}D'| = |D|^2/|\Delta|$$

In particular if $g = (\theta_1, \theta_2, \dots \theta_k)$ then

$$|\Sigma_{\mathbf{T}}| \geq 1/|\Delta|.$$

Remark 4 In the above theorem we have assumed that the dimension of the parameter is the same as that of the vector of functions that we want to estimate. This need not be the case. Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_m)$ and $\mathbf{g} = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \dots g_k(\boldsymbol{\theta}))$ and let $D = (d_{ij})$ be a $k \times m$ matrix of $\frac{\partial}{\partial \theta_j} g_i(\boldsymbol{\theta}), i = 1, 2, \dots k, j = 1, 2, \dots m$. Then the theorem states that $\Sigma_{\mathbf{T}} - D\Delta^{-1}D'$ is n.n.d.$\forall \theta \in \Theta \subset \mathcal{R}^m$. The proof is exactly similar in case $k = m$. The case $k < m$ is usually of interest which corresponds to say for example, estimating $(\theta_1, \theta_2, \dots \theta_k)$ only while $(\theta_{k+1}, \theta_{k+2}, \dots \theta_m)$ acts as a nuisance parameter. In this case the theorem states that $\Sigma_{\mathbf{T}} - \Delta^{11}$ is n.n.d. where Δ^{-1} is partitioned as

$$\Delta^{-1} = \begin{pmatrix} \Delta^{11} & \Delta^{12} \\ \Delta^{21} & \Delta^{22} \end{pmatrix}.$$

Example Let $X_1, X_2, \dots X_n$ be i.i.d. $N(\mu, \sigma^2)$ variables. Writing $\boldsymbol{\theta} = (\theta_1, \theta_2)$ where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. the p.d.f. of X_i is given by

$$f_{\boldsymbol{\theta}}(x) = \frac{1}{\theta_2\sqrt{2\pi}} exp\left\{\frac{-1}{2\theta_2}(x - \theta_1)^2\right\}, -\infty < x < \infty, -\infty < \theta_1 < \infty, \theta_2 > 0.$$

We know that $T_1 = \bar{X}$ and $T_2 = S^2$ are unbiased estimators of θ_1 and θ_2 respectively where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.
Since $\bar{X} \sim N(\theta_1, \frac{\theta_2}{n})$ and $\frac{(n-1)S^2}{\theta_2} \sim \chi^2_{n-1}$ we get
$V_{\boldsymbol{\theta}}(T_1) = \frac{\theta_2}{n}$ and $Var_{\boldsymbol{\theta}}(T_2) = \frac{2\theta^2}{n-1}$.
As $\bar{X}$ and S^2 are independent we get $Cov_{\boldsymbol{\theta}}(T_1, T_2) = 0$. Hence the variance-covariance matrix of T is given by

$$\Sigma_{\mathbf{T}} = \begin{pmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n-1} \end{pmatrix}.$$

The elements of the score vector $\mathbf{S} = (S_1, S_2)$ are given by

$$S_1 = \frac{\partial lnf}{\partial \theta_1} = \frac{(x - \theta_1)}{\theta_2}$$

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

and

$$S_2 = \frac{\partial lnf}{\partial \theta_2} = -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2}$$

Thus

$$E_{\boldsymbol{\theta}}\left(\frac{\partial lnf}{\partial \theta_1}\right)^2 = \frac{1}{\theta_2^2}E_{\boldsymbol{\theta}}(X-\theta_1)^2 = \frac{1}{\theta_2},$$

$$\begin{aligned} E_{\boldsymbol{\theta}}\left(\frac{\partial lnf}{\partial \theta_2}\right)^2 &= \frac{1}{4\theta_2^2} + \frac{1}{4\theta_2^4}E_{\boldsymbol{\theta}}(X-\theta_1)^4 - \frac{1}{2\theta_2^3}E_{\boldsymbol{\theta}}(X-\theta_1)^2 \\ &= \frac{1}{4\theta_2^2} + \frac{1}{4\theta_2^4}3\theta_2^2 - \frac{1}{2\theta_2^3}\theta_2 \\ &= \frac{1}{2\theta_2^2} \end{aligned}$$

and

$$\begin{aligned} E_{\boldsymbol{\theta}}\left[\left(\frac{\partial lnf}{\partial \theta_1}\right)\left(\frac{\partial lnf}{\partial \theta_2}\right)\right] &= -\frac{1}{2\theta_2^2}E_{\boldsymbol{\theta}}(X-\theta_1) + \frac{1}{2\theta_2^3}E_{\boldsymbol{\theta}}(X-\theta_1)^3 \\ &= 0 \end{aligned}$$

As $X_i's$ are i.i.d.

$$\delta_{11} = \frac{n}{\theta_2},\ \delta_{22} = \frac{n}{2\theta_2^2} \text{ and } \delta_{12} = 0.$$

Hence the information matrix is given by

$$\Delta = \begin{pmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{pmatrix}.$$

As in this case $D = I_2$, we get

$$\Sigma_{\mathbf{T}} - \Delta^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{2\theta_2^2}{n(n-1)} \end{pmatrix}.$$

Note : Since T_1 and T_2 are respectively the UMVUE's of θ_1 and θ_2 from the above example we see that the the variance of T_1 attains the Cramer-Rao lower bound but variance of T_2 does not.

Statistical Inference I
Interval Estimation

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Introduction

So far we have discussed the point estimation of a parameter, or more precisely, point estimation of several real valued parametric functions. Such point estimates are quite useful, yet they leave something to be desired. In case of continuous distributions the probability that the point estimator actually equaled the value of the parameter being estimated is zero. Hence, it seems desirable that a point estimate should be accompanied by some measure of the possible error of estimate. For instance, a point estimate must be accompanied by some interval about the point estimate together with some measure of assurance that the true value of the parameter lies within the interval. Instead of making the inference of estimating the true value of the parameter to be a point, we might make the inference that the true value of the parameter is contained in some interval. This is called the problem of interval estimation.

Confidence Interval

Let $X_1, X_2, \ldots X_n$ be a random sample from a population with p.m.f. or p.d.f. $f_\theta(x)$ where $\theta \in \Theta(\subset \mathcal{R})$ is unknown. Let T_1 and T_2 be two statistics satisfying $T_1 \leq T_2$ for which $P_\theta(T_1 \leq \theta \leq T_2) = 1 - \alpha$, where $\alpha \in (0, 1)$ does not depend on θ. Thus the interval $[T_1, T_2]$ will include the parameter θ with probability $(1 - \alpha)$. This means that if a large number of random samples of same size be taken, and the intervals as mentioned above be determined on the basis of these samples, the $100(1 - \alpha)\%$ of these intervals will include the true value of the parameter θ. The random interval (T_1, T_2) is called a $100(1 - \alpha)\%$ confidence interval for θ. The quantity $(1 - \alpha)$ is called the confidence level and T_1 and T_2 are called the lower and upper confidence

limits,respectively,for θ.

Example 1. Let $X_1, \ldots, X_n \sim N(\mu, 1)$. Since we know, $\bar{X} \sim N(\mu, \frac{1}{\sqrt{n}})$. Therefore,

$$P_\mu \left[\tau_{\alpha/2} \leq \sqrt{n}\left(\bar{X} - \mu\right) \leq \tau_{\alpha/2}\right] = 1 - \alpha \ \forall \mu \in R.$$

Hence by interchanging sides,

$$P_\mu \left[\bar{X} - \frac{\tau_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{\tau_{\alpha/2}}{\sqrt{n}}\right] = 1 - \alpha \ \forall \mu \in R.$$

Hence $[\bar{x} - \frac{\tau_{\alpha/2}}{\sqrt{n}}, \bar{x} + \frac{\tau_{\alpha/2}}{\sqrt{n}}]$ is a $100(1-\alpha)\%$ confidence interval of θ.

Determination of confidence interval using a Pivotal quantity

Let $\mathbf{X} = (X_1, X_2, \ldots X_n)$ be a random sample from a population with p.m.f or p.d.f $f_\theta(x)$. Let $Q(x, \theta)$ be a real valued function of X and θ is called a pivot if its distribution is completely known under $\theta \in \Theta$. i.e. the Q has a distribution that does not depend on θ.

Since the distribution of Q is independent of θ we can find two numbers q_1 and q_2 $(q_1 < q_2)$ such that,

$$P_\theta \left[q_1 \leq Q(x, \theta) \leq q_2\right] \geq 1 - \alpha \ \forall \theta \in \Theta$$

$$\Leftrightarrow P_\theta \left[t_1(x) \leq \theta \leq t_2(x)\right] \geq 1 - \alpha \ \forall \theta \in \Theta.$$

Hence $[t_1(x), t_2(x)]$ is a confidence interval for θ at confidence level $1 - \alpha$.

Shortest confidence interval based on a pivot:

For any fixed $(1-\alpha)$ there are many possible pairs of numbers q_1 and q_2 that can be selected so that $P(q_1 < \theta < q_2) = 1 - \alpha$. Different pairs of q_1 and q_2 will produce t_1 and t_2. We should want to select the pair of q_1 and q_2 that will make t_1 and t_2 close together in some sense. For instance, if $t_2(\mathbf{x}) - t_1(\mathbf{x})$ which is the length of the confidence interval, is not random, then we might select select that pair of q_1 and q_2 that makes the length of the interval smallest; or if the length of the confidence interval is random, then we might select that pair q_1 and q_2 that makes the average length of the interval smallest.

Thus a shortest length confidence(SLCI) interval is the confidence interval for which $t_2(x) - t_1(x)$ is minimized and a shortest expected length confidence interval(SELCI) is the confidence interval for which $E_\theta \{T_2(\mathbf{X}) - T_1(\mathbf{X})\}$ is minimized $\forall\, \theta \in \Theta$.

Example 2. Let $\mathbf{X} = X_1, \ldots, X_n \sim N(\mu, 1)$. To find shortest length confidence interval for μ, based on the pivot $Q(\mathbf{X}, \mu) = \sqrt{n}(\bar{X} - \mu)$. We know that $Q(\mathbf{X}, \mu) \sim N(0, 1)$ for any μ.

$$\begin{aligned}
P_\mu \left[q_1 \leq Q(\mathbf{X}, \mu) \leq q_2\right] &= 1 - \alpha \\
\text{or, } P_\mu \left[q_1 \leq \sqrt{n}(\bar{X} - \mu) \leq q_2\right] &= 1 - \alpha \\
\text{or, } P_{\mu,\sigma} \left[\bar{X} - \frac{q_1}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{q_2}{\sqrt{n}}\right] &= 1 - \alpha.
\end{aligned}$$

Therefore, $t_1(x) = \bar{X} - \frac{q_1}{\sqrt{n}}$ and $t_2(x) = \bar{X} + \frac{q_2}{\sqrt{n}}$, hence $T_2(x) - T_1(x) = (q_2 - q_1)\frac{1}{\sqrt{n}}$. is not random variable. Our objective is to minimize $q_2 - q_1$ subject to,

$$P[q_1 \leq Z \leq q_2] = 1 - \alpha, Z \sim N(0, 1).$$

Let us define,

$$\psi(q_1, q_2) = (q_2 - q_1) + \lambda \left[F_Z(q_2) - F_Z(q_1) - (1 - \alpha)\right]$$

Hence, by differentiating $\psi(q_1, q_2)$ w.r.t. q_1 and q_2 we get the following set of equations, and solving for 0 gives us the shortest length confidence interval.

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = -1 - \lambda f_Z(q_1) = 0$$

and,

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = 1 + \lambda f_Z(q_2) = 0$$

$$\Leftrightarrow f_Z(q_1) = f_Z(q_2)$$

Therefore, $q_2 = -q_1$ and hence $q_2 = \tau_{\frac{\alpha}{2}}$, where $\tau_{\frac{\alpha}{2}}$ denotes the upper $100\frac{\alpha}{2}$ point of $N(0,1)$ distribution.

$$\text{SLCI} : \left[\bar{x} - \frac{\tau_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{x} + \frac{\tau_{\frac{\alpha}{2}}}{\sqrt{n}}\right].$$

<u>Example 3.</u> Let $\mathbf{X} = X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. To find shortest length confidence interval for μ, based on the pivot $Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\bar{X}-\mu)}{s}$. We know that $Q(\mathbf{X}, \mu) \sim t_{n-1}$ for any μ, σ.

$$\begin{aligned}
P_{\mu,\sigma}\left[q_1 \leq Q(\mathbf{X}, \mu) \leq q_2\right] &= 1-\alpha \\
\text{or, } P_{\mu,\sigma}\left[q_1 \leq \frac{\sqrt{n}(\bar{X}-\mu)}{s} \leq q_2\right] &= 1-\alpha \\
\text{or, } P_{\mu,\sigma}\left[\bar{X} - \frac{q_1 s}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{q_2 s}{\sqrt{n}}\right] &= 1-\alpha.
\end{aligned}$$

Therefore, $t_1(x) = \bar{X} - \frac{q_1 s}{\sqrt{n}}$ and $t_2(x) = \bar{X} + \frac{q_2 s}{\sqrt{n}}$, hence $t_2(x) - t_1(x) = (q_2 - q_1)\frac{s}{\sqrt{n}}$. Our objective is to minimize $q_2 - q_1$ subject to,

$$P[q_1 \leq t_{n-1} \leq q_2] = 1 - \alpha.$$

Let us define,

$$\psi(q_1, q_2) = (q_2 - q_1) + \lambda\left[F_{t_{n-1}}(q_2) - F_{t_{n-1}}(q_1) - (1-\alpha)\right]$$

Hence, by differentiating $\psi(q_1, q_2)$ w.r.t. q_1 and q_2 we get the following set of equations, and solving for 0 gives us the shortest length confidence interval.

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = -1 - \lambda f_{t_{n-1}}(q_1) = 0$$

and,

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = 1 + \lambda f_{t_{n-1}}(q_2) = 0$$

$$\Leftrightarrow f_{t_{n-1}}(q_1) = f_{t_{n-1}}(q_2)$$

Therefore, $q_2 = -q_1$ and hence $q_2 = t_{\frac{\alpha}{2},n-1}$ and,

$$\text{SLCI} : \left[\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right].$$

Example 4. Let $\mathbf{X} = X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. To find shortest length confidence interval for σ^2, based on the pivot $Q(\mathbf{X}, \sigma^2) = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{\sigma^2}$. We know that $Q(\mathbf{X}, \sigma^2) \sim \chi^2_{n-1}$ for any μ, σ^2.

$$\begin{aligned}
P_{\mu,\sigma}\left[q_1 \le Q(\mathbf{X}, \mu) \le q_2\right] &= 1-\alpha \\
\text{or, } P_{\mu,\sigma}\left[q_1 \le \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{\sigma^2} \le q_2\right] &= 1-\alpha \\
\text{or, } P_{\mu,\sigma}\left[\frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{q_2} \le \sigma^2 \le \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{q_1}\right] &= 1-\alpha.
\end{aligned}$$

Hence, $t_2(x) - t_1(x) = \sum_{i=1}^n \left(X_i - \bar{X}\right)^2 \left(\frac{1}{q_1} - \frac{1}{q_2}\right)$. To minimize $\frac{1}{q_1} - \frac{1}{q_2}$ subject to,

$$P[q_1 \le \chi^2_{n-1} \le q_2] = 1-\alpha.$$

Let us define,

$$\psi(q_1, q_2) = \left(\frac{1}{q_1} - \frac{1}{q_2}\right) + \lambda\left[F_{\chi^2_{n-1}}(q_2) - F_{\chi^2_{n-1}}(q_1) - (1-\alpha)\right]$$

Hence, by differentiating $\psi(q_1, q_2)$ w.r.t. q_1 and q_2 we get the following set of equations, and solving for 0 gives us the shortest length confidence interval.

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = -\frac{1}{q_1^2} - \lambda f_{\chi^2_{n-1}}(q_1) = 0$$

and,

$$\frac{\partial}{\partial q_1}\psi(q_1, q_2) = \frac{1}{q_2^2} + \lambda f_{\chi^2_{n-1}}(q_2) = 0$$

Solving for the above two equations simultaneously, we get

$$\begin{aligned}
\Leftrightarrow q_1^2 f_{\chi^2_{n-1}}(q_1) &= q_2^2 f_{\chi^2_{n-1}}(q_2) \\
\Rightarrow \exp^{-q_1/2} q_1^{(n+1)/2} &= \exp^{-q_2/2} q_2^{(n+1)/2}.
\end{aligned}$$

Example 5. Let $\mathbf{X} = X_1, \ldots, X_n \sim R(0, \theta)$. To find shortest length confidence interval for θ, based on the pivot $Q(X_{(n)}, \theta) = \frac{X_{(n)}}{\theta}$. Note that, the probability density of $W = X_{(n)}$ is given,

$$p_\theta(w) = \frac{n}{\theta^n} w^{n-1},\ 0 < w < \theta.$$

If we consider the following transformation $Y = \frac{W}{\theta}$,

$$p(y) = ny^{n-1},\ 0 < y < 1.$$

$$\begin{aligned}
\int_{q_1}^{q_2} ny^{n-1} dy = P\left[q_1 < Y < q_2\right] &= 1 - \alpha \\
P_\theta\left[q_1 < \frac{X_{(n)}}{\theta} < q_2\right] &= 1 - \alpha \\
P_\theta\left[\frac{1}{q_2} < \frac{\theta}{X_{(n)}} < \frac{1}{q_1}\right] &= 1 - \alpha \\
P_\theta\left[\frac{X_{(n)}}{q_2} < \theta < \frac{X_{(n)}}{q_1}\right] &= 1 - \alpha
\end{aligned}$$

Hence, $t_2(x) - t_1(x) = X_{(n)}\left(\frac{1}{q_1} - \frac{1}{q_2}\right)$. To minimize $\frac{1}{q_1} - \frac{1}{q_2}$ subject to,

$$\int_{q_1}^{q_2} ny^{n-1} dy = (q_2^n - q_1^n) = 1 - \alpha. \tag{1}$$

Now $(1-\alpha)^n < q_2 \le 1$ and

$$\frac{dL}{dq_2} = \frac{-1}{q_1^2}\frac{dq_1}{dq_2} + \frac{1}{q_2^2}. \tag{2}$$

Again from (1) we get

$$nq_2^{n-1} - nq_1^{n-1}\frac{dq_1}{dq_2} = 0$$

$$\Rightarrow \frac{dq_1}{dq_2} = \frac{q_2^{n-1}}{q_1^{n-1}}.$$

Hence from (2) we get

$$\begin{aligned}
\frac{dL}{dq_2} &= \frac{-1}{q_1^2}\frac{q_2^{n-1}}{q_1^{n-1}} + \frac{1}{q_2^2} \\
&= \frac{-q_2^{n-1} + q_1^{n-1}}{q_1^{n+1}q_2^2} < 0
\end{aligned}$$

Hence, L is a decreasing function of q_2. So L is minimum if q_2 is maximum and the maximum value occurs at $q_2 = 1$ and hence $q_1 = \alpha^{1/n}$.Hence $100(1-\alpha)\%$ shortest length confidence interval for θ is $\left(X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}}\right)$.

<u>Example 6.</u> Let $(\mathbf{X}, \mathbf{Y}) = \{(X_1, Y_1); \ldots; (X_n, Y_n)\} \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. To find shortest length confidence interval for ρ. Let $\theta = \mu_1/\mu_2$ be the parameter of interest with $\mu_2 \neq 0$. Define, $Z_i(\theta) = Y_i - \theta X_i \left(\sim N(0, \sigma_2^2 - 2\theta\rho + \theta^2\sigma_1^2)\right)$. Let the sample variance for $Z_i(\theta)$,

$$S^2(\theta) = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i(\theta) - \bar{\mathbf{Z}}(\theta)\right) = S_2^2 - 2\theta S_{12} + \theta^2 S_1^2,$$

where, S_i and S_{12} are sample variances and covariance of X_i and Y_i. Now from example 3 we know that $\sqrt{n}\bar{\mathbf{Z}}(\theta)/S(\theta) \sim t_{n-1}$ and therefore, is a pivotal quantity $Q((\mathbf{X}, \mathbf{Y}), \theta) = \sqrt{n}\bar{\mathbf{Z}}(\theta)/S(\theta)$. Hence,

$$P_\theta\left[q_1 \leq Q((\mathbf{X}, \mathbf{Y}), \theta) \leq q_2\right] = 1 - \alpha.$$

We know $-q_1 = q_2 = t_{n-1,\alpha/2}$ therefore by symmetry,

$$\begin{aligned} \text{or, } P_\theta\left[\frac{n\bar{\mathbf{Z}}^2(\theta)}{S^2(\theta)} \leq t^2_{n-1,\alpha/2}\right] &= 1 - \alpha \\ \text{or, } P_\theta\left[S_2^2 - 2\theta S_{12} + \theta^2 S_1^2 \leq \frac{n\bar{\mathbf{Z}}^2(\theta)}{t^2_{n-1,\alpha/2}}\right] &= 1 - \alpha. \end{aligned}$$

Note that, $t^2_{n-1,\alpha}\left(S_2^2 - 2\theta S_{12} + \theta^2 S_1^2\right) = n\bar{\mathbf{Z}}^2(\theta)$ defines a parabola in θ. Depending on the roots of θ the interval can be a finite interval, the complement of a finite interval of the whole real line.

Remark

Like point estimation the problem of interval estimation is twofold. First there is a problem of finding interval estimators,and, second, there is a problem of determining good, or optimum, interval estimators. In this lecture we have discussed a method of finding a good confidence interval by minimizing the expected length of the interval. The other criteria of getting an optimum confidence will be discussed at the end of the lecture on testing of hypothesis.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Introduction to Testing of Hypothesis
Module- 21

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In Modules 1 - 19 we have discussed the theory point estimation and in Module-20 we have introduced the basic idea of interval estimation. Now we are going to introduce the notions of testing of hypothesis. In point estimation some characteristic or feature of the population in which we are interested may be completely unknown to us and we may like to make a guess about the characteristic entirely on the basis of a random sample drawn from the population whereas in interval estimation we try to put forward an interval with the hope that the interval would contain the unknown parameter with some preassigned probability. In testing of hypothesis, some information regarding the characteristic or feature of the population may be available to us and we may like to know whether the information is tenable (or can be accepted) in the light of the random sample drawn from the population.

Historical Perspective The first published test of a statistical hypothesis was by John Arbuthnott (Kendall and Plackett,1977) in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls. He considered data on the annual number of christenings in London classified by sex for the 82 years 1629-1710. Equating christenings with live births, his intention was to test the hypothesis H: births represent independent trials with a constant probability 0.5 of a male child being born. If H holds, he would say that chance governs the birth of a child. He used the original data only to check whether each year was a male year (i.e. having a majority of male births) or not. As it turned out, all the 82 years considered by Arbuthnott were male years.

Using on the theory of Bernoulli trials he observed that if H holds then, irrespective of the number of births in different years, the probability of all 82 years considered being male years does not exceed $(\frac{1}{2})^{82}$. As this is exceedingly small, Arbuthnott rejected H and concluded that 'it is Art, not chance that governs' the birth of a child. He argued that this was a proof of divine providence, since some of the boys will be soldiers they have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition that the male birth ratio is 0.5, related to data by regarding these as outcomes of stochastic variables, the calculation that the data would be unlikely if the proposition were true, and a further conclusion interpreting the falsity of the proposition.The above idea was first used by Karl Pearson in 1900 for proposing the well-known Chi-squared test of goodness of fit, and later by Fisher in proposing various tests of significance. An unified approach for testing statistical hypothesis was developed by Jerzy Neyman and Egon Pearson in 1928.

Notations and preliminaries

Let θ be the labelling parameter of the distribution of a random variable X and Θ be the parametric space. As in the case of parametric estimation problem θ can be either real valued or vector valued. Instead of $f_\theta(x)$ here we use $p(x,\theta)$ as the p.m.f. or p.d.f. of the random variable X and the family of the distribution is denoted as $P = \{p(x,\theta); \theta \in \Theta\}$. The objective here is to know whether $p(x,\theta) \in P_H \subset P$ on the basis of $X = x$. So our conjecture is H which is our null hypothesis. Whenever we fix H we have a corresponding hypothesis called alternative hypothesis, often denoted by H_A. Specifically we write H to mean the null hypothesis that $H : p(x,\theta) \in P_H = \{p(x,\theta); \theta \in \Theta_H\}$, which is equivalent to $H : \theta \in \Theta_H$. On other hand H_A is represented as $H_A : p(x,\theta) \in P_{H_A} = \{p(x,\theta); \theta \in \Theta_{H_A}\}$, which is equivalent to $H_A : \theta \in \Theta_{H_A}$. Further note that, $\Theta_H, \Theta_{H_A} \subset \Theta$ and $\Theta_H \cap \Theta_{H_A} = \phi$. Depending on the nature of a the hypothesis we can classify it as a simple or a com-

posite hypothesis.

Simple and Composite hypothesis

A statistical hypothesis H is called simple if it completely specifies the distribution of the random variable X; otherwise it is called composite. If Θ_H is a single point set then H is simple ; otherwise it is composite.

Example Let $X \sim N(\theta, 1)$. If we want to test $H : \theta = 0$ against $H_A : \theta > 0$, then H is simple but H_A is composite.

Example Let $X \sim N(\mu, \sigma^2)$ where both μ and σ are unknown. If we want to test $H : \mu = 0$ against $H_A : \mu > 0$, then both H and H_A are composite because σ is unknown.

Test of a hypothesis

A test of a statistical hypothesis H is a rule or procedure by which one can accept or reject H. Two types of test are used in practice one is called a non-randomized test and the other is called a randomized test.

Non-randomized test (Deterministic test)

Let $\mathcal{X}$ be the sample space. A Non-randomized test is a rule by which one chooses a region $\omega(\subset \mathcal{X})$ before the start of the experiment such that whenever $X = x$ is observed, H is rejected if $x \in \omega$; otherwise it is accepted. The region ω is called the critical region and $\mathcal{X} - \omega$ is called the acceptance region of the test. A Non-randomized test is also called a deterministic test because here the critical region is found completely before the testing and no probabilistic statement is attached with the region.

Test statistic

In practice, associated with any critical region ω, it is possible to find a statistic

ePathshala पाठशाला A Gateway to All Post Graduate Courses

$T = T(X)$ such that,

$$[x \in \omega] \Leftrightarrow \{x : T(x) \geq C_u\} : \text{ Upper-tail test (one-sided alternative)}$$

or,

$$[x \in \omega] \Leftrightarrow \{x : T(x) \leq C_l\} : \text{ Lower-tail test (one-sided alternative)}$$

or,

$$[x \in \omega] \Leftrightarrow \{x : T(x) \leq C_l \text{ or } T(x) \geq C_u\} : \text{ Two-tailed test (Two-sided alternative) .}$$

Such a statistic $T(X)$ is called a test statistic. It is used in specifying the critical region ω.

Example Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $N(\theta, 1)$ population. For testing $H : \theta = 0$ against $H_A : \theta > 0$ suppose one rejects H if $\sum_{i=1}^{n} X_i > c$ where c is a constant. Then $\sum_{i=1}^{n} X_i$ is the test statistic for testing H and the critical region of the test is $\omega = \{(X_1, X_2, \ldots X_n) : \sum_{i=1}^{n} X_i > c\}$.

Two types of errors

In a non-randomized testing problem there will be either correct or incorrect decision. Correct decision can arise for either of the case when we accept H if it is true of reject H when it is false. On the other hand incorrect decision can be either if we reject H if it is false and accept H when it is false. The measures used to quantify the two scenarios of incorrect decisions are called type I error and type II error. Type I error is the one, if we reject H when it is true, while type II error is the one, if we accept H when it is false. For example. in statistical quality control Type I error occurs when a lot of good quality is being rejected and Type II error occurs when a lot of bad quality is being accepted. The two probabilities of two types of errors are as follows

$$P(TypeIerror) = P(X \in \omega | \theta \in \Theta_H) \text{ and } P(TypeIIerror) = P(X \in \mathcal{X} - \omega | \theta \in \Theta_{H_A})$$

To study the behavior of two error probabilities let us define a function $P_\omega(\theta) = P\{X \in \omega|\theta\}$.The function defined by $P_\omega(\theta)$ on Θ is called power function of the test with critical region ω. Then

$$P(Type I error) = \beta_\omega(\theta), \theta \in \Theta_H \text{and} P(Type II error) = 1 - \beta_\omega(\theta), \theta \in \Theta_{H_A}$$

An ideal test procedure would be one for which $\beta_\omega(\theta), \theta \in \Theta_H$ and $1-\beta_\omega(\theta), \theta \in \Theta_{H_A}$ are minimized simultaneously by selecting ω in a proper way. However, even in the simplest problem where $\Theta_H = \{\theta_0\}$ and $\Theta_{H_A} = \{\theta_1\}$, we can not select ω such that both $\beta_\omega(\theta_0)$ and $1 - \beta_\omega(\theta_1)$ are minimized simultaneously. This follows from the fact that $\beta_\omega(\theta_0)$ can be reduced only by removing sample points from ω i.e. shrinking ω, but this results ion reducing $\beta_\omega(\theta_1)$ which increases $1 - \beta_\omega(\theta_1)$. On the other hand $1 - \beta_\omega(\theta_1)$ can be reduced only by increasing $\beta_\omega(\theta_1)$ or by enlarging the set ω which increases $\beta_\omega(\theta_0)$.

Example Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $N(\theta, 1)$ population. For testing $H : \theta = \theta_0$ against $H_A : \theta = \theta_1(> \theta_0)$ suppose one rejects H if $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i > c$ where c is a constant. Then the error probabilities are given by

$$P(Type I error) = P_{\theta_0}(\bar{X} > c) = 1 - \Phi(\sqrt{n}(c - \theta_0))$$

$$P(Type II error) = P_{\theta_1}(\bar{X} < c) = \Phi(\sqrt{n}(c - \theta_1))$$

If we want to reduce $P(Type I error)$ then we have to increase c and hence $P(Type II error)$ will increase. On the other hand, if we want to reduce $P(Type II error)$ then we have to decrease c. As a result $P(Type I error)$ will increase.

Power of a test

The quantity $\beta_\omega(\theta)$, $\theta \in \Theta_{H_A}$ is called power of the test ω at different θ. Hence power of a test represents the probability of a correct decision and it is equal to $1 - P(Type II error)$.

Let us consider some examples.

Example Let $X \sim Bin(5, p)$. For testing $H : p = \frac{1}{2}$ against $p = \frac{3}{4}$ the following test

is used:

$$\omega = \{x : x \geq 3\}$$

Then

$$P(TypeIerror) = P(X \geq 3 \mid X \sim Bin(5, \frac{1}{2})) = \sum_{x=4}^{5} \binom{5}{x} \left(\frac{1}{2}\right)^5 = \frac{3}{16}$$

$$P(TypeIIerror) = P(X \leq 3 \mid X \sim Bin(5, \frac{1}{4})) = 1 - \sum_{x=4}^{5} \binom{5}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{5-x} = \frac{47}{128}$$

The power of the test $= 1 - \frac{47}{128} = \frac{81}{128}$.

Example Let X be a random variable with p.d.f.

$$p(x, \theta) = \theta e^{-\theta x}, x > 0, \theta > 0.$$

For testing $H : \theta = 2$ against $H_A : \theta = 1$ suppose H is rejected if $x \geq 1$. Then

$$P(TypeIerror) = P_H(X \geq 1) = \int_1^\infty 2e^{-2x} dx = \frac{1}{e^2}$$

$$P(TypeIIerror) = P_{H_A}(X \leq 1) = \int_0^1 e^{-x} dx = 1 - \frac{1}{e}$$

The power of the test $=\frac{1}{e}$.

Example Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $N(\theta, 1)$ population. For testing $H : \theta = \theta_0$ against $H_A : \theta = \theta_1 (> \theta_0)$ suppose one rejects H if $\bar{X} > \theta_0 + \frac{\tau_\alpha}{\sqrt{n}}$, where τ_α is the upper $100\alpha\%$ point of $N(0, 1)$ distribution. The power function of the test is

$$\beta_\omega(\theta) = P_\theta \left(\bar{X} > \theta_0 + \frac{\tau_\alpha}{\sqrt{n}} \right) = 1 - \Phi(\tau_\alpha + \sqrt{n}(\theta_0))$$

The power of the test at $\theta = \theta_1 (> \theta_0)$ is equal to $\beta_\omega(\theta_1)$.

Statistical Inference I
Idea of Test function
Module- 22

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In the previous module we have seen that when the sample size is fixed it is not possible to minimize both the error probabilities at the same time. A test that minimizes the probability of type-I error, in fact maximizes the probability of type-II error. To overcome this problem, Neyman-Pearson suggested that one may select a suitable upper bound for the probability of type-I error and consider only those critical regions for which the probability of type-I error does not exceed the selected upper bound. Secondly, among all these critical regions select the one for which probability of type-II error is the smallest. This upper bound is called the level of significance of the test.

Level of Significance A test with critical region ω is called a level $\alpha \in [0,1)$ if $P(X \in \omega | \theta \in \Theta_H) \leq \alpha$ where $\alpha \in [0,1)$ is a preassigned quantity and it is called the level of significance of the test.

Note A level α test is also a level α_0 test ($\alpha < \alpha_0 < 1.$). Here α and α_0 are upper bounds of the set $\{P(X \in \omega | \theta \in \Theta_H)\}$. It is quite natural to look for the least upper bound of the set $\{P(X \in \omega | \theta \in \Theta_H)\}$. This is called the size of the test.

Size of a test A test with critical region ω is called a level $\alpha \in [0,1)$ if $SupP(X \in \omega | \theta \in \Theta_H) = \alpha$.

Note A size α test is also a level α test. A test with level α is easily attainable whereas it is not easy to attain the exact size α when the underlying distribution is discrete.

Example Let θ be the probability of getting a head in single toss of a coin. For testing $H : \theta = \frac{1}{2}$ against $H_A : \theta > \frac{1}{2}$ suppose the coin is tossed 5 times and H is

rejected if $X > c$ where X denotes the number of heads obtained in 5 tosses and $c > 0$ is a constant.Here $P_H(X > 3) = 0.1875, P_H(X > 4) = 0.03125$. If we take $c = 4$ then the test is a level 0.05 test but not a size 0.05 test.
To attain the exact size we use the randomized test.

Randomized test (Probabilistic test): It consist of determining a real-valued function $\varphi(x)$:
$mathcal X \Leftrightarrow [0,1]\,(\Leftrightarrow 0 \le \varphi(x) \le 1 \forall x \in \chi)$ such that, whenever $X = x$ is observed a test is rejected or accepted according as success or failure is occurred, by performing a bernoulli trial with probability of success equal to $\phi(x)$. Here φ is called test or critical function. Here it can be noted that ω is a special choice of φ where,

$$\omega = \{x; \varphi(x) = 1\}$$

$$\chi - \omega = \{x; \varphi(x) = 0\}$$

In general, a randomized test is of the following form:

$$\begin{aligned} \varphi(x) &= 1 \text{ if } x \in E_1 \\ &= \gamma \text{ if } x \in E_2 \\ &= 0 \text{ if } x \in E_3 \end{aligned}$$

where $E_i's$ are mutually disjoint subsets of $\mathcal{X}$ with $E_1 \bigcup E_2 \bigcup E_3 = \mathcal{X}$. The constants $c > 0$ and $\gamma \in (0,1)$ are so chosen that the test is of size α.
In terms of the test statistic $T(X)$ a two tailed randomized test is given by

$$\begin{aligned} \varphi(x) &= 1 \text{ if } T(x) \le c_l or T(x) \ge c_u \\ &= 0 \text{ if } c_l \le T(x) \le c_u \end{aligned}$$

with randomization at $T(X) = c_l$ or $T(X) = c_u$.

Example In the above example if we want to get an exact size 0.05 test we set $c = 4$. Then to determine γ we use the equation

$$P_H(X > 4) + \gamma P_H(X = 4) = 0.05$$

$$\Rightarrow \gamma = 0.12.$$

Power function

$$\varphi: \text{ Probability of rejection of } H \text{ when } X = x \text{ is observed.}$$

Hence, unconditional probability of rejection of H is

$$\begin{aligned} P_\varphi(\theta) &= \int_{\mathcal{X}} \varphi(x) p(\mathbf{x}, \theta) d\mathbf{x} \\ &= E_\theta(\varphi(X)), \theta \in \Theta \end{aligned} \tag{0.1}$$

Therefore,

$$P(\text{Type I error}) = E_\theta(\varphi(X)), \theta \in \Theta_H$$

$$\Longleftrightarrow \text{ Size of } \varphi = \sup_{\theta \in \Theta_H} \{E_\theta(\varphi(X)\}.$$

And when H is simple, $\Theta_H = \theta_0$, size is $E_{\theta_0}(\varphi(X))$.

$$\begin{aligned} P(\text{ Type II error }) &= E_\theta(1 - \varphi(X)), \theta \in \Theta_{H_A} \\ &= 1 - E_\theta(\varphi(X)) \\ &= 1 - \text{ power of the } \varphi \text{ at } \theta, \theta \in \Theta_{H_A}. \end{aligned} \tag{0.2}$$

In general, the power function of φ is defined by

$$P_\varphi(\theta) = E_\theta(\varphi(X)), \theta \in \Theta.$$

φ is a level α test if,

$$E_\theta(\varphi(X)) \leq \alpha, \ \ \forall \theta \in \Theta_H$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

$$\Longleftrightarrow sup_{\theta\in\Theta_H} E_\theta\varphi(X) \leq \alpha,$$

where $0 < \alpha < 1$.

Result. Let φ_1 and φ_2 be two level α tests for (H, H_A). Then any convex combination of φ_1 and φ_2 is also a level α test for (H, H_A).

Proof. Take $0 \leq \lambda \leq 1$, and set

$$\varphi(x) = \lambda\varphi_1(x) + (1-\lambda)\varphi_2(x).$$

Here, $\varphi : \chi \longrightarrow [0,1]$. Moreover,

$$E_\theta(\varphi(X)) = \lambda E_\theta(\varphi_1(X)) + (1-\lambda)E_\theta(\varphi_2(X)) \leq \lambda\alpha + (1-\lambda)\alpha, \forall\theta \in \Theta_H,$$

$$\Longleftrightarrow E_\theta(\varphi(X)) \leq \alpha\forall\theta \in \Theta_H.$$

p-value. Instead of fixing level of significance α one can obtain the level actually obtained by a test. This is defined by,

$$l^+(x) = P_\theta\left\{T(X) \geq T(x)\right\} : \text{ Upper-tail test based on T(x)}$$

$$l^-(x) = P_\theta\left\{T(X) le T(x)\right\} : \text{ Lower-tail test based on T(x)}$$

$$l(x) = 2\min(l^+(x), l^-(x)) : \text{ Both tailed test based on T(x).}$$

$\theta \in \Theta_H$, provided the distribution of $T(X)$ remains same over the null hypothetical set Θ_H. If we assume that the distribution of $T(x)$ is continuous, we have,

$$l^-(x) + l^+(x) = 1.$$

Moreover,

$$l^-(x), l^+(x) \sim R(0,1) \text{ under } \theta \in \Theta_H.$$

[Actually,

$$l^+(x) = 1 - F(T(X)), F \text{ being the d.f. of } T$$

$$l^-(x) = F(T(X))$$

]

Smaller value of l^+(or l^-) gives stronger evidence in favor of the rejection of H (or against H). Further it can be shown that,

$$l(X) \sim R(0,1) \text{ under H.}$$

Again if the distribution of $T(X)$ is discrete, we modify our level actually attained values by,

$$l^+(x) = P_\theta(T(X) > T(x)) + \frac{1}{2}P_\theta(T(X) = T(x))$$
$$l^-(x) = P_\theta(T(X) < T(x)) + \frac{1}{2}P_\theta(T(X) = T(x))$$

for $\theta \in \Theta_H$. Here $l^+(x)$ and $l^-(x)$ are approximately $R(0,1)$ where under H. It is easy to check that under continuous setup,

$$l(x) = 2\min(l^+(x), l^-(x)) \sim R(0,1) \text{ whenever } \theta \in \Theta_H.$$

Some Examples

Example 1. $X_1, X_2, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$

$$H : \mu = 0 \text{ vs. } H_A : \mu > 0$$

$$(i)\sigma = 1 \text{ (known) Test statistic } = T(X) = \frac{\sqrt{n}\bar{X}}{\sigma} \sim N(0,1) \text{ under } H : \mu = 0$$

$$(ii)\sigma \text{ (unknown) Test statistic } = T(X) = \frac{\sqrt{n}\bar{X}}{\sigma} \sim t_{n-1} \text{ under } H : \mu = 0\text{and}$$

Suppose we want to test

$$H : \sigma = 1 \text{ vs. } H_A : \sigma > 1$$

$$(iii)\mu \text{ (known) Test statistic } = T(X) = \sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n \text{ under H}$$

$$(iii)\mu \text{ (unknown) Test statistic } = T(X) = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1} \text{ under H}$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Subject - Statistics
Paper - Statistical Inference I
Module - The Neyman-Pearson Fundamental Lemma [1]

Most Powerful Tests

Let, $\mathscr{P}$ be a family of probability measures $\big($defined on a measurable space $(\Omega, \mathscr{A})\big)$, indexed by a parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d, d \in \{1, 2, 3, \ldots\}$.

$$\mathscr{P} := \big\{\mathbb{P}_{\boldsymbol{\theta}} \colon \mathbb{P}_{\boldsymbol{\theta}} \text{ is a probability measure on } (\Omega, \mathscr{A}),\, \boldsymbol{\theta} \in \Theta\big\}$$

Consider the problem of testing $H_0 \colon \boldsymbol{\theta} \in \Theta_0$ against $H_1 \colon \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \cup \Theta_1 \subseteq \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$. Let, $\mathscr{M}_\alpha$ be the class of all level α tests for (H_0, H_1), $\alpha \in \,]0, 1[$.

$$\mathscr{M}_\alpha := \big\{\varphi \colon \varphi \text{ is a test function, } \mathbb{E}_{\boldsymbol{\theta} \in \Theta_0}\big[\varphi(\boldsymbol{X})\big] \leqslant \alpha\big\}$$

Let, $\boldsymbol{X} := \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}^\top$ be an n-tuple of IID random variables (a random sample), taking value in a set $\mathfrak{X}$ $(\subseteq \mathbb{R}^n)$, with joint density (with respect to some σ-finite measure μ) $p(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$, corresponding to some $\mathbb{P}_{\boldsymbol{\theta}} \in \mathscr{P}$.

Definition: A level α test φ_0 $(\in \mathscr{M}_\alpha)$ is said to be *Most Powerful* (MP) for testing $H_0 \colon \boldsymbol{\theta} \in \Theta_0$ against $H_1 \colon \boldsymbol{\theta} = \boldsymbol{\theta}_1$ $(\notin \Theta_0)$ if

$$\mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi_0(\boldsymbol{X})\big] \geqslant \mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi(\boldsymbol{X})\big], \forall\, \varphi \in \mathscr{M}_\alpha.$$

The Neyman-Pearson Fundamental Lemma

Theorem: [*Existence*] For testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1\, (\neq \boldsymbol{\theta}_0)$ at some level α, there exists a test φ_0 given by

$$\varphi_0(\boldsymbol{x}) = \begin{cases} 0, & \text{if } p(\boldsymbol{x}; \boldsymbol{\theta}_1) < k \cdot p(\boldsymbol{x}; \boldsymbol{\theta}_0) \\ 1, & \text{if } p(\boldsymbol{x}; \boldsymbol{\theta}_1) > k \cdot p(\boldsymbol{x}; \boldsymbol{\theta}_0); \end{cases} \tag{1}$$

where k $(\geqslant 0)$ and φ_0 on the boundary $\big\{\boldsymbol{x} \colon p(\boldsymbol{x}; \boldsymbol{\theta}_1) = k \cdot p(\boldsymbol{x}; \boldsymbol{\theta}_0)\big\}$ are so determined that

$$\mathbb{E}_{\boldsymbol{\theta}_0}\big[\varphi_0(\boldsymbol{X})\big] = \alpha. \tag{2}$$

[1] Co-Ordinator: Dr. Shirsendu Mukherjee, Asutosh College, Kolkata

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

[*Sufficiency*] Furthermore, if a test satisfies (1) and (2), for some constant k ($\geqslant 0$), then it is MP for (H_0, H_1) at level α.

[*Necessity*] Again, if φ_0 is MP at level α for (H_0, H_1), then for some k ($\geqslant 0$), it satisfies (1) almost everywhere. It also satisfies (2), unless there exists a test of size less than α and with power 1.

Proof:

In this module, we shall prove the *Existence* and *Sufficiency* parts of the theorem; the proof of *Necessity*, requires some prerequisites, and will be taken up in the next module.

[*Existence*] We shall, first, show that a test satisfying (1) and (2) actually exists.

Define, $Y := p(\boldsymbol{X}; \boldsymbol{\theta}_1) \div p(\boldsymbol{X}; \boldsymbol{\theta}_0)$.

As $\mathbb{P}_{\boldsymbol{\theta}_0}\big\{p(\boldsymbol{X}; \boldsymbol{\theta}_0) > 0\big\} = 1$, so $\mathbb{P}_{\boldsymbol{\theta}_0}\{0 \leqslant Y < \infty\} = 1$. In other words, Y is well defined under H_0.

Let

$$\begin{aligned} \alpha(y) &:= \mathbb{P}_{\boldsymbol{\theta}_0}\big\{p(\boldsymbol{X}; \boldsymbol{\theta}_1) > y \cdot p(\boldsymbol{X}; \boldsymbol{\theta}_0)\big\} \\ &= \mathbb{P}_{\boldsymbol{\theta}_0}\big\{p(\boldsymbol{X}; \boldsymbol{\theta}_1) > y \cdot p(\boldsymbol{X}; \boldsymbol{\theta}_0), p(\boldsymbol{X}; \boldsymbol{\theta}_0) > 0\big\} \\ &= \mathbb{P}_{\boldsymbol{\theta}_0}\{Y > y\} \;=:\; 1 - G_{\boldsymbol{\theta}_0}(y), \end{aligned}$$

where, $G_{\boldsymbol{\theta}_0}(\cdot)$ is the CDF of Y, under H_0.

$G_{\boldsymbol{\theta}_0}(\cdot)$ is non-decreasing and right-continuous, and hence, $\alpha(\cdot)$ is non-increasing and right-continuous. Moreover, $\alpha(0-) = 1 - G_{\boldsymbol{\theta}_0}(0-) = 1$ and $\alpha(\infty) = 1 - G_{\boldsymbol{\theta}_0}(\infty) = 0$. Hence, given α ($\in]0,1[$), we can find k ($\geqslant 0$) such that

$$\alpha(k-) \geqslant \alpha \geqslant \alpha(k). \tag{3}$$

Here, two cases can arise

- If k is a point of continuity, we have $\alpha(k) = \alpha(k-) = \alpha$. Then, the test given by (1) satisfies (2), regardless of the choice of φ_0 at the boundary.

- If k is a point of discontinuity, we set φ_0 at the boundary as

$$\gamma := \frac{\alpha - \alpha(k)}{\alpha(k-) - \alpha(k)}. \tag{4}$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}_0}\big[\varphi_0(\boldsymbol{X})\big] &= \mathbb{P}_{\boldsymbol{\theta}_0}\big\{p(\boldsymbol{X};\boldsymbol{\theta}_1)>k\cdot p(\boldsymbol{X};\boldsymbol{\theta}_0)\big\} + \gamma\cdot\mathbb{P}_{\boldsymbol{\theta}_0}\big\{p(\boldsymbol{X};\boldsymbol{\theta}_1)=k\cdot p(\boldsymbol{X};\boldsymbol{\theta}_0)\big\} \\
&= \alpha(k) + \frac{\alpha-\alpha(k)}{\alpha(k-)-\alpha(k)}\cdot\big[\alpha(k-)-\alpha(k)\big] \;=\; \alpha.
\end{aligned}$$

So, a test of the form (1), satisfying (2), exists. [Q.E.D.]

[*Sufficiency*] Our objective, here, is to show that any test of the form (1), satisfying (2), is most powerful.

Pick any $\varphi \in \mathscr{M}_\alpha$. Then, for any $\boldsymbol{x} \in \mathfrak{X}$, consider the product

$$\big[\varphi_0(\boldsymbol{x}) - \varphi(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\theta_1) - k\cdot p(\boldsymbol{x};\theta_0)\big].$$

By (1), when $p(\boldsymbol{x};\theta_1) > k\cdot p(\boldsymbol{x};\theta_0)$, then $\varphi_0(\boldsymbol{x}) = 1 \geqslant \varphi(\boldsymbol{x})$. Similarly, when $p(\boldsymbol{x};\theta_1) < k\cdot p(\boldsymbol{x};\theta_0)$, then $\varphi_0(\boldsymbol{x}) = 0 \leqslant \varphi(\boldsymbol{x})$. In both cases, the product $\big[\varphi_0(\boldsymbol{x}) - \varphi(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\theta_1) - k\cdot p(\boldsymbol{x};\theta_0)\big]$ is non-negative. Of course, when $p(\boldsymbol{x};\theta_1) = k\cdot p(\boldsymbol{x};\theta_0)$, the product is zero, and hence non-negative.

$$\begin{aligned}
\therefore \quad & \big[\varphi_0(\boldsymbol{x}) - \varphi(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\theta_1) - k\cdot p(\boldsymbol{x};\theta_0)\big] \;\geqslant\; 0 \qquad\qquad (5) \\
\implies \quad & \int_{\mathfrak{X}}\big[\varphi_0(\boldsymbol{x}) - \varphi(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\theta_1) - k\cdot p(\boldsymbol{x};\theta_0)\big]\mathsf{d}\boldsymbol{x} \;\geqslant\; 0 \\
\iff \quad & \mathbb{E}_{\boldsymbol{\theta}_1}\left[\varphi_0(\boldsymbol{X})\right] - \mathbb{E}_{\boldsymbol{\theta}_1}\left[\varphi(\boldsymbol{X})\right] \;\geqslant\; k\cdot\big[\mathbb{E}_{\boldsymbol{\theta}_0}\left[\varphi_0(\boldsymbol{X})\right] - \mathbb{E}_{\boldsymbol{\theta}_0}\left[\varphi(\boldsymbol{X})\right]\big] \\
& \qquad\qquad\qquad\qquad\qquad\quad = k\cdot\big[\alpha - \mathbb{E}_{\boldsymbol{\theta}_0}\left[\varphi(\boldsymbol{X})\right]\big] \;\geqslant\; 0. \\
& \qquad\qquad\qquad\qquad \big[\because\; \mathbb{E}_{\boldsymbol{\theta}_0}\left[\varphi(\boldsymbol{X})\right] \;\leqslant\; \alpha.\big]
\end{aligned}$$

$$\therefore\; \mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi_0(\boldsymbol{X})\big] \;\geqslant\; \mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi(\boldsymbol{X})\big],\; \forall\,\varphi \in \mathscr{M}_\alpha.$$

Thus, φ_0 is MP in its class. [Q.E.D.]

Subject - Statistics
Paper - Statistical Inference I
Module - The Neyman-Pearson Fundamental Lemma [1]

Some Measure Theoretic Preliminaries

Let Ω be a set, and $\mathscr{A}$ be a σ-algebra of subsets of Ω. The pair $(\Omega, \mathscr{A})$ is called a measurable space. $\boldsymbol{X} : \Omega \longrightarrow \mathfrak{X}$ $(\subseteq \mathbb{R}^n)$ is an $\mathscr{A}/\mathscr{B}$-measurable map, where $\mathscr{B}$ is a σ-algebra of subsets of $\mathfrak{X}$, usually the σ-algebra of Borel subsets of $\mathfrak{X}$. Clearly, $(\mathfrak{X}, \mathscr{B})$ defines another measurable space. We define a measure μ on $(\mathfrak{X}, \mathscr{B})$ by

$$\mu(\mathbb{B}) := \int_{\mathbb{B}} \mathsf{d}\boldsymbol{x},$$

where $\mathbb{B}$ is a set in $\mathscr{B}$. The triple $(\mathfrak{X}, \mathscr{B}, \mu)$ is called a measure space.

For a function $g(\cdot)$, defined on $\mathfrak{X}$, we define the integral of $g(\cdot)$ over a set $\mathbb{B}$ $(\in \mathscr{B})$ as

$$\int_{\mathbb{B}} g(\boldsymbol{x})\mathsf{d}\boldsymbol{x} = \int_{\mathfrak{X}} \mathbf{1}_{\mathbb{B}}(\boldsymbol{x}) g(\boldsymbol{x})\mathsf{d}\boldsymbol{x},$$

where $\mathbf{1}_{\mathbb{B}}(\boldsymbol{x}) := \begin{cases} 1, & \text{if } \boldsymbol{x} \in \mathbb{B} \\ 0, & \text{if } \boldsymbol{x} \notin \mathbb{B} \end{cases}$ is the indicator of $\mathbb{B}$.

If $g(\cdot)$ is non-negative on $\mathbb{B}$,

$$\int_{\mathbb{B}} g(\boldsymbol{x})\mathsf{d}\boldsymbol{x} = 0 \Longrightarrow \begin{cases} \text{either,} & g(\boldsymbol{x}) = 0 \text{ on } \mathbb{B} \\ \text{or,} & \mu(\mathbb{B}) = 0. \end{cases}$$

We say that $g(\boldsymbol{x}) = 0$, almost everywhere (a.e.) on $\mathbb{B}$, or more specifically, $g(\boldsymbol{x}) = 0$, μ-a.e. on $\mathbb{B}$. That is, $g(\boldsymbol{x}) = 0, \forall\, \boldsymbol{x} \in \mathbb{B} \backslash N$, where $\mu(N) = 0$ with $N \in \mathscr{B}$.

- If $g(\cdot)$ is non-negative a.e. on $\mathbb{B}$ and $\int_{\mathbb{B}} g(\boldsymbol{x})\mathsf{d}\boldsymbol{x} = 0$, then $g(\boldsymbol{x}) = 0$ a.e. on $\mathbb{B}$.
- If $g(\cdot)$ is positive a.e. on $\mathbb{B}$ and $\int_{\mathbb{B}} g(\boldsymbol{x})\mathsf{d}\boldsymbol{x} = 0$, then $\mu(\mathbb{B}) = 0$.

[1] Co-Ordinator: Dr. Shirsendu Mukherjee, Asutosh College, Kolkata

epathshala A Gateway to All Post Graduate Courses

The Necessity Part of the Neyman-Pearson Lemma

In the last module, we stated the Neyman-Pearson fundamental lemma, proved the *Existence* and *Sufficiency* parts of the lemma, but stopped short of proving *Necessity*. In this module, having briefly digressed to discuss some pre-requisite measure theory, we return to the lemma.

Before preceeding to the proof, we recapitulate the *Necessity* part of the lemma.

If a test φ_0 is MP at level α for testing $H_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1\colon \boldsymbol{\theta} = \boldsymbol{\theta}_1$ ($\neq \boldsymbol{\theta}_0$), then for some non-negative real number k, it satisfies

$$\varphi_0(\boldsymbol{x}) = \begin{cases} 0, & \text{if } p(\boldsymbol{x};\theta_1) < k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0) \\ 1, & \text{if } p(\boldsymbol{x};\theta_1) > k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0) \end{cases} \tag{1}$$

almost everywhere. It also satisfies

$$\mathbb{E}_{\boldsymbol{\theta}_0}\big[\varphi_0(\boldsymbol{X})\big] = \alpha, \tag{2}$$

unless there exists a test of size less than α and with power 1.

Proof:

[*Necessity*] Here, we shall show that any MP test at level α for (H_0, H_1) is, necessarily, of the form (1), except on a set of μ-measure 0.

As before, φ_0 is a test satisfying (1) and (2). Also, let $\varphi^\star$ be MP at level α for (H_0, H_1). We have [cf. (5) and the argument leading upto it, in the proof of *Sufficiency*],

$$\big[\varphi_0(\boldsymbol{x}) - \varphi^\star(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\boldsymbol{\theta}_1) - k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0)\big] \geqslant 0,\ \forall\, \boldsymbol{x} \in \mathfrak{X}.$$

Consider the sets:

$$A_1 := \big\{\boldsymbol{x}\colon \varphi_0(\boldsymbol{x}) \neq \varphi^\star(\boldsymbol{x})\big\} \text{ and } A_2 := \big\{\boldsymbol{x}\colon p(\boldsymbol{x};\boldsymbol{\theta}_1) \neq k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0)\big\}.$$

Then,

$$\begin{aligned} &\int_{\mathfrak{X}} \big[\varphi_0(\boldsymbol{x}) - \varphi^\star(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\boldsymbol{\theta}_1) - k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0)\big]\mathsf{d}\boldsymbol{x} \\ = &\int_{A_1 \cap A_2} \big[\varphi_0(\boldsymbol{x}) - \varphi^\star(\boldsymbol{x})\big]\big[p(\boldsymbol{x};\boldsymbol{\theta}_1) - k \cdot p(\boldsymbol{x};\boldsymbol{\theta}_0)\big]\mathsf{d}\boldsymbol{x}. \end{aligned}$$

If $\mu(A_1 \cap A_2) > 0$, then $\mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi_0(\boldsymbol{X}) - \varphi^\star(\boldsymbol{X})\big] > 0$, i.e., φ_0 is more powerful than $\varphi^\star$, which is a contradiction. So, to get $\mathbb{E}_{\boldsymbol{\theta}_1}\big[\varphi_0(\boldsymbol{X}) - \varphi^\star(\boldsymbol{X})\big] = 0$, we must have $\mu(A_1 \cap A_2) = 0$, i.e., $\varphi_0(\boldsymbol{x}) = \varphi^\star(\boldsymbol{x})$ μ-a.e.

In other words, MP tests are unique (of the form (1)), except on the set $\big\{\boldsymbol{x} : p(\boldsymbol{x}; \boldsymbol{\theta}_1) = k \cdot p(\boldsymbol{x}; \boldsymbol{\theta}_0)\big\}$. [Q.E.D.]

Having shown how to 'build' a most powerful test, we prove, next, an important property of theirs. In any reasonable test, we would want the power to exceed the size — a condition, technically, called unbiasedness [not to be confused with the notion of an unbiased estimator].

Theorem: Most powerful tests are unbiased.

Proof: Let φ_0 be a most powerful test at level α. Take $\varphi(\boldsymbol{x}) \equiv \alpha$. Obviously, $\varphi \in \mathscr{M}_\alpha$, and $\mathbb{E}_{\boldsymbol{\theta}}\big[\varphi(\boldsymbol{X})\big] = \alpha$, $\forall\, \boldsymbol{\theta} \in \Theta$. As φ_0 is MP at level α, so

$$\text{Power}(\varphi_0) = \mathbb{E}_{\boldsymbol{\theta}_1}[\varphi_0(\boldsymbol{X})] \geqslant \mathbb{E}_{\boldsymbol{\theta}_1}[\varphi(\boldsymbol{X})] = \alpha = \text{Size}(\varphi_0). \text{ [Q.E.D.]}$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Hypothesis Testing in Uniform [0,θ] - I

The Set-up

Data:

$X_1, X_2, , X_n \overset{\text{iid}}{\sim}$ Unif $[0, \theta], \theta > 0$. The X_j's take value in $\mathscr{X} = \{\mathbf{x} := (x_1, x_2, \ldots, x_n) : x_j \geqslant 0, \forall\, j\}$, which we call our *sample space*. θ is the unknown parameter here. Our inferential procedures will focus on testing hypotheses on θ.

The p.d.f. of each X_j, is of the form

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}, & \text{if } x \leqslant \theta \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The joint p.d.f. of $X_1, X_2, \ldots, X_n$, is

$$p_\theta(\mathbf{x}) = \prod_{j=1}^{n} f_\theta(x_j) = \begin{cases} \frac{1}{\theta^n}, & \text{if } x_{(n)} \leqslant \theta \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $x_{(n)}$ is the maximum of the x_j's. For $n = 2$, we may visualize the sample space $\mathscr{X}$ as in figure 1

The Testing Problem

Suppose we are interested in testing

$$H_0 : \theta = \theta_0 \text{ (known) against } H_1 : \theta > \theta_0.$$

We start with

$$H_0 : \theta = \theta_0 \text{ versus } H_1' : \theta = \theta_1 (> \theta_0).$$

Our aim is to find a most powerful (MP) test for the problem of testing H_0 against H_1'. Figure 2 explains the hypotheses H_0 and H_1'.

Illustration of the Problem

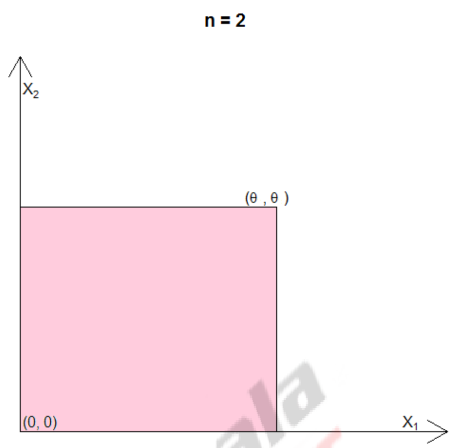


Figure 1: Visualization for $n = 2$

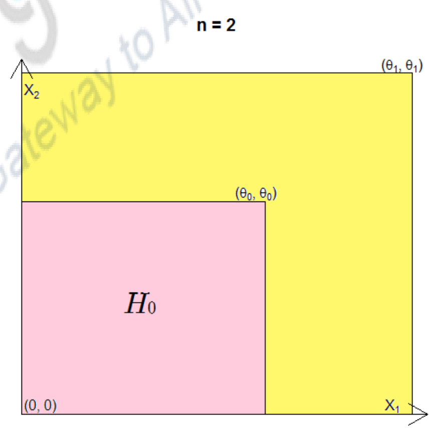


Figure 2: H_0 vs. H_1'

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

The region in pink is the sample space, under H_0. Mathematically, the pink (P) and yellow (Y) zones can be identified as

$$P := \{\mathbf{x} : x_{(n)} \leqslant \theta_0\},$$

and,

$$Y := \{\mathbf{x} : \theta_0 < x_{(n)} \leqslant \theta_1\}.$$

Derivation of MP Tests

$$p_{\theta_1}(\mathbf{x}) = \begin{cases} \frac{1}{\theta_1^n}, & \text{if } x_{(n)} \leqslant \theta_1 \\ 0, & \text{if } x_{(n)} > \theta_1. \end{cases} \tag{3}$$

$$k \cdot p_{\theta_0}(\mathbf{x}) = \begin{cases} \frac{k}{\theta_0^n}, & \text{if } x_{(n)} \leqslant \theta_0 \\ 0, & \text{if } x_{(n)} > \theta_0. \end{cases} \tag{4}$$

Take $k(\geqslant 0)$ such that

$$\frac{1}{\theta_1^n} = \frac{k}{\theta_0^n}$$

$$\Leftrightarrow \quad k = \left(\frac{\theta_0}{\theta_1}\right)^n.$$

$p_{\theta_1}(\mathbf{x}) > k \cdot p_{\theta_0}(\mathbf{x}) \Leftrightarrow p_{\theta_1}(\mathbf{x}) = \frac{1}{\theta_1^n}$ and $k \cdot p_{\theta_0}(\mathbf{x}) = 0$,i.e., $\theta_0 < x_{(n)} \leqslant \theta_1$. The region in violet depicts what the inequalities above say. $p_{\theta_1}(\mathbf{x}) < k \cdot p_{\theta_0}(\mathbf{x}) \Leftrightarrow p_{\theta_1}(\mathbf{x}) = 0$ and $k \cdot p_{\theta_0}(\mathbf{x}) = \frac{k}{\theta_0^n}$, i.e., $x_{(n)} > \theta_1$ and $x_{(n)} \leqslant \theta_0$. The regions in violet depict what the inequalities above say. Clearly, both inequalities cannot, simultaneously, hold.

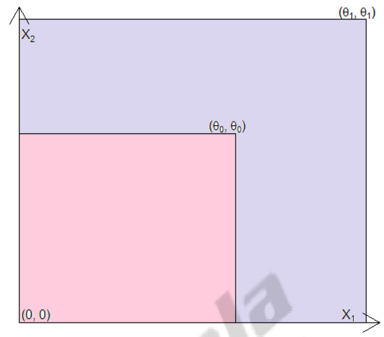
An MP Test: By the Neyman-Pearson lemma, an MP level α test for testing H_0 against $H_1' : \theta = \theta_1$, is of the form

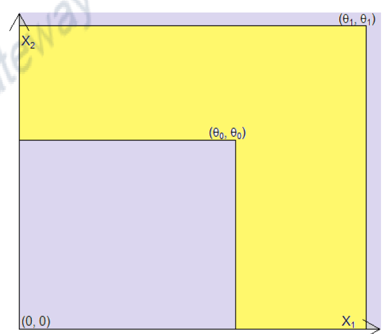
$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \text{if } \theta_0 < x_{(n)} \leqslant \theta_1 \\ \text{any value in } [0,1], & \text{if } x_{(n)} > \theta_1 \text{ or } x_{(n)} \leqslant \theta_0, \end{cases}$$

such that

$$E_{\theta_0}\big[\phi_0(\mathbf{X})\big] = \alpha.$$

We will look at some special tests of this form. The test ϕ_0 is free to take any value when $x_{(n)} > \theta_1$ or $x_{(n)} \leqslant \theta_0$, as long as $E_{\theta_0}\big[\phi_0(\mathbf{X})\big] = \alpha$.





epathshala पाठशाला A Gateway to All Post Graduate Courses

So, to make the test free of θ_1, i.e., to make it UMP for testing H_0 against $H_1\colon \theta > \theta_0$, we write

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \text{if } x_{(n)} > \theta_0 \\ \text{any value in } [0,1], & \text{if } x_{(n)} \leqslant \theta_0, \end{cases} \tag{5}$$

such that

$$E_{\theta_0}\big[\phi_0(\mathbf{X})\big] = \alpha. \tag{6}$$

<u>A Randomized MP Test:</u> As a randomized UMP at level α test for testing H_0 against $H_1\colon \theta > \theta_0$, we modify (1) into

$$\phi_{01}(\mathbf{x}) = \begin{cases} 1, & \text{if } x_{(n)} > \theta_0 \\ \alpha, & \text{if } x_{(n)} \leqslant \theta_0. \end{cases} \tag{7}$$

Of course, $E_{\theta_0}\big[\phi_{01}(\mathbf{X})\big] = \alpha \cdot P_{\theta_0}\big\{X_{(n)} \leqslant \theta_0\big\} = \alpha.$

<u>Non-randomized MP Tests:</u> An alternative to a randomized test is a non-randomized level α UMP test.

We have to choose a subset A_{θ_0} (depending only on θ_0) of the region $\mathscr{X}_{\theta_0} := \{\mathbf{x}\colon x_{(n)} \leqslant \theta_0\}$ (on which to reject H_0), such that the size restriction $E_{\theta_0}\big[\phi_0(\mathbf{X})\big] = \alpha$ holds.

Define,

$$\phi_{02}(\mathbf{x}) = \begin{cases} 1, & \text{if } x_{(n)} > \theta_0 \\ 1, & \text{if } \mathbf{x} \in A_{\theta_0} \\ 0, & \text{if } \mathbf{x} \in \mathscr{X}_{\theta_0} - A_{\theta_0}, \end{cases} \tag{8}$$

where A_{θ_0} is such that

$$P_{\theta_0}\{\mathbf{X} \in A_{\theta_0}\} = \alpha. \tag{9}$$

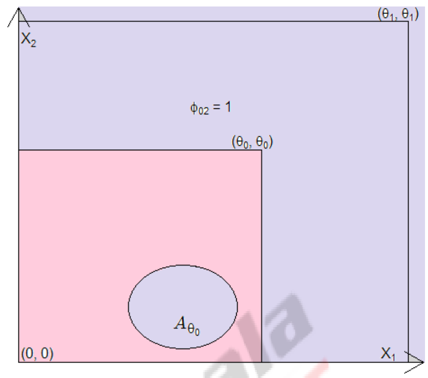
The test ϕ_{02} of the form (4), satisfying (5), can be diagramatically elucidated thus:

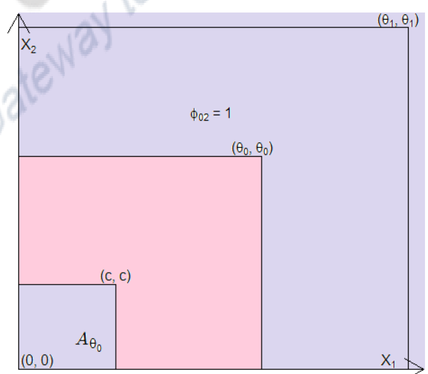
At level α, ϕ_{02} rejects H_0 in the violet regions, and accepts in the pink region. As ϕ_{02} is independent of the choice of θ_1, as long as $\theta_1 > \theta_0$, this test is, also, UMP for testing $H_0\colon \theta = \theta_0$ against $H_1\colon \theta > \theta_0$.

<u>Choice of A_{θ_0}:</u> Take $A_{\theta_0} = \{\mathbf{x}\colon x_{(n)} \leqslant c\}$, for some constant c $(< \theta_0)$.

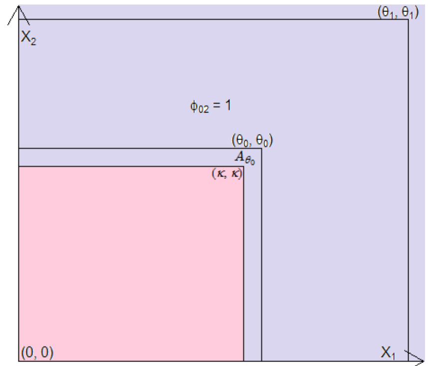
c satisfies

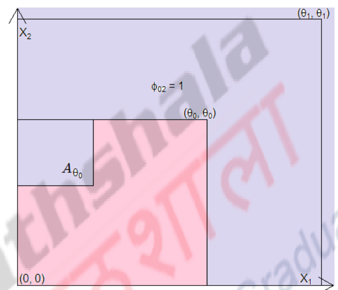
$$\begin{aligned} & P_{\theta_0}\{X_{(n)} \leqslant c\} = \alpha \\ \Rightarrow \quad & \left(\tfrac{c}{\theta_0}\right)^n = \alpha \\ \Leftrightarrow \quad & c = \theta_0 \cdot \sqrt[n]{\alpha}. \end{aligned} \tag{10}$$





e-PG Pathshala पाठशाला A Gateway to All Post Graduate Courses

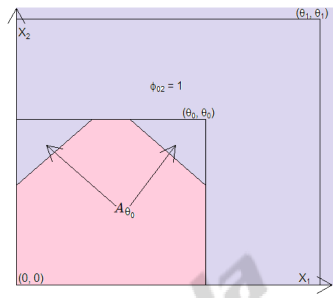




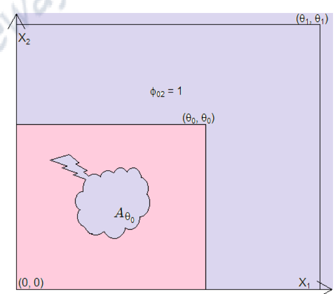
Take $A_{\theta_0} = \{\mathbf{x}\colon x_{(n)} > \kappa\}$, for some constant κ ($< \theta_0$).
κ satisfies

$$\begin{aligned} & P_{\theta_0}\{X_{(n)} > \kappa\} = \alpha \\ \Rightarrow \quad & 1 - \left(\frac{\kappa}{\theta_0}\right)^n = \alpha \\ \Leftrightarrow \quad & \kappa = \theta_0 \cdot \sqrt[n]{1-\alpha}. \end{aligned} \tag{11}$$

Other Choices of A_{θ_0}: Any subset of $\mathscr{X}_{\theta_0} = \{\mathbf{x}\colon x_{(n)} \leqslant \theta_0\}$, satisfying (5), will serve as a potential A_{θ_0}. A few other choices are given below:



ePathshala पाठशाला A Gateway to All Post Graduate Courses



Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Hypothesis Testing in Uniform $[0,\theta]$ - II

Power Function

Before computing the power functions for the tests, we note, For $a \leqslant \theta$,

$$P_\theta\{X_{(n)} \leqslant a\} = \left(\frac{a}{\theta}\right)^n, \tag{1}$$

$$P_\theta\{X_{(n)} > a\} = 1 - \left(\frac{a}{\theta}\right)^n; \tag{2}$$

While, for $b > \theta$,

$$P_\theta\{X_{(n)} \leqslant b\} = 1, \tag{3}$$

$$P_\theta\{X_{(n)} > b\} = 0. \tag{4}$$

Power Function for ϕ_{01}: Consider the UMP level α randomized test ϕ_{01}, given by

$$\phi_{01}(\mathbf{x}) = \begin{cases} \alpha, & \text{if } x_{(n)} \leqslant \theta_0 \\ 1, & \text{if } x_{(n)} > \theta_0. \end{cases} \tag{5}$$

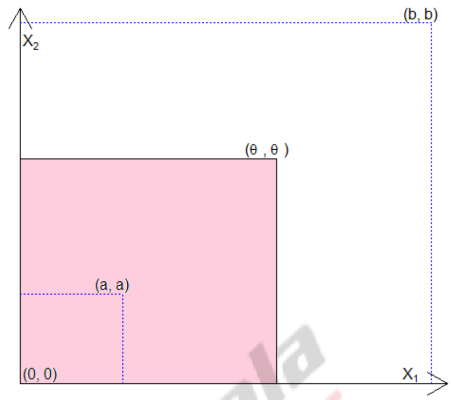
The power function of this test is

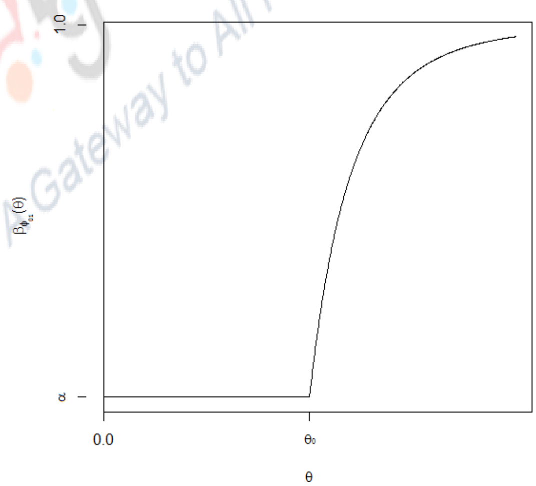
$$\beta_{\phi_{01}}(\theta) = E_\theta\Big[\phi_{01}(\mathbf{X})\Big] = \Big[\alpha \times P_\theta\big\{X_{(n)} \leqslant \theta_0\big\}\Big] + \Big[1 \times P_\theta\big\{X_{(n)} > \theta_0\big\}\Big] \tag{6}$$

$$= \begin{cases} \alpha, & \text{if } \theta \leqslant \theta_0 \\ \left[\alpha \cdot \left(\frac{\theta_0}{\theta}\right)^n\right] + \left[1 - \left(\frac{\theta_0}{\theta}\right)^n\right], & \text{if } \theta > \theta_0. \end{cases}$$

We have,

$$\beta_{\phi_{01}}(\theta) = \begin{cases} \alpha, & \text{if } \theta \leqslant \theta_0 \\ 1 - (1-\alpha)\left(\frac{\theta_0}{\theta}\right)^n, & \text{if } \theta > \theta_0. \end{cases} \tag{6}$$





epathshala पाठशाला
A Gateway to All Post Graduate Courses

Clearly, $\beta_{\phi_{01}}(\theta) \nearrow \theta$, and $\beta_{\phi_{01}}(\theta) \leqslant \alpha, \forall\, \theta \leqslant \theta_0$. So, ϕ_{01} is UMP at level α, for testing

$$H_0' : \theta \leqslant \theta_0 \text{ against } H_1 : \theta > \theta_0,$$

as well.

Power Function for Tests of the form ϕ_{02}: We shall, now, derive the power function(s) for the test(s) given by (4) and (5):

$$\phi_{02}(\mathbf{x}) = \begin{cases} 1, & \text{if } x_{(n)} > \theta_0 \\ 1, & \text{if } \mathbf{x} \in A_{\theta_0} \\ 0, & \text{if } \mathbf{x} \in \mathscr{X}_{\theta_0} - A_{\theta_0}, \end{cases}$$

where A_{θ_0} is such that $P_{\theta_0}\{\mathbf{X} \in A_{\theta_0}\} = \alpha$.

$$P_{\theta_0}\{\mathbf{X} \in A_{\theta_0}\} = \alpha.$$

Such a test has been shown to be UMP for testing,

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta > \theta_0.$$

We will look at the power function of the special forms of this test, already discussed; viz., when $A_{\theta_0} = \{\mathbf{x} : x_{(n)} \leqslant c\}$, and when $A_{\theta_0} = \{\mathbf{x} : x_{(n)} > \kappa\}$.

Case when $A_{\theta_0} = \{\mathbf{x} \colon x_{(n)} > \kappa\, (< \theta_0)\}$:

We have seen that the size restriction forces $\kappa = \theta_0 \cdot \sqrt[n]{1-\alpha}$. Then the test becomes

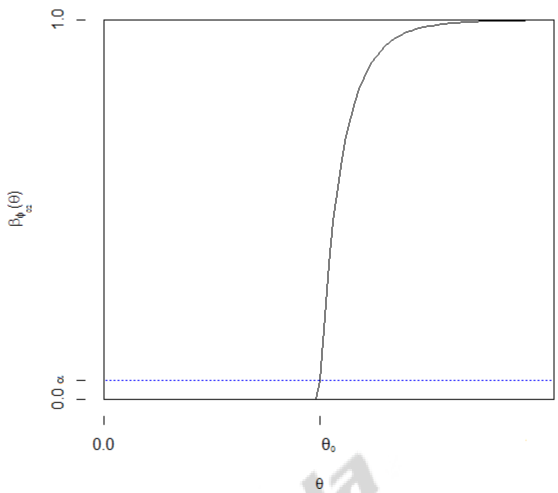
$$\phi_{02}(\mathbf{x}) = \begin{cases} 0, & \text{if } x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{1-\alpha} \\ 1, & \text{if } x_{(n)} > \theta_0 \cdot \sqrt[n]{1-\alpha}. \end{cases} \tag{8}$$

The power function of this test is,

$$\begin{aligned} \beta_{\phi_{02}}(\theta) &= P_\theta\{X_{(n)} > \theta_0 \cdot \sqrt[n]{1-\alpha}\} \\ &= \begin{cases} 0, & \text{if } \theta \leqslant \theta_0 \cdot \sqrt[n]{1-\alpha} \\ 1 - (1-\alpha) \cdot \left(\frac{\theta_0}{\theta}\right)^n, & \text{if } \theta > \theta_0 \cdot \sqrt[n]{1-\alpha}. \end{cases} \end{aligned} \tag{9}$$

Clearly, $\beta_{\phi_{02}}(\theta) \nearrow \theta$, and $\beta_{\phi_{02}}(\theta) \leqslant \alpha, \forall\, \theta \leqslant \theta_0$. So, ϕ_{02} (as given by (8)) is UMP at level α, for testing

$$H_0' : \theta \leqslant \theta_0 \text{ against } H_1 : \theta > \theta_0.$$



Case when $A_{\theta_0} = \{\mathbf{x} : x_{(n)} \leqslant c(< \theta_0)\}$*:*

We have seen that the size restriction forces $c = \theta_0 \cdot \sqrt[n]{\alpha}$. Then the test becomes

$$\phi_{02}(\mathbf{x}) = \begin{cases} 1, & \text{if } \{x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha}\} \text{ or } \{x_{(n)} > \theta_0\} \\ 0, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < x_{(n)} \leqslant \theta_0. \end{cases} \tag{10}$$

The power function of this test is

$$\begin{aligned} \beta_{\phi_{02}}(\theta) &= P_\theta\{X_{(n)} > \theta_0\} + P_\theta\{X_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha}\} \\ &= \begin{cases} 1, & \text{if } \theta \leqslant \theta_0 \cdot \sqrt[n]{\alpha} \\ \alpha \cdot \left(\frac{\theta_0}{\theta}\right)^n, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < \theta \leqslant \theta_0 \\ 1 - (1-\alpha) \cdot \left(\frac{\theta_0}{\theta}\right)^n, & \text{if } \theta > \theta_0. \end{cases} \end{aligned} \tag{11}$$

$\beta_{\phi_{02}}(\theta)$ (as given by (11)) is U-shaped, with a minimum value of α at θ_0. Though the test is UMP for (H_0, H_1), it is not UMP for (H_0', H_1).

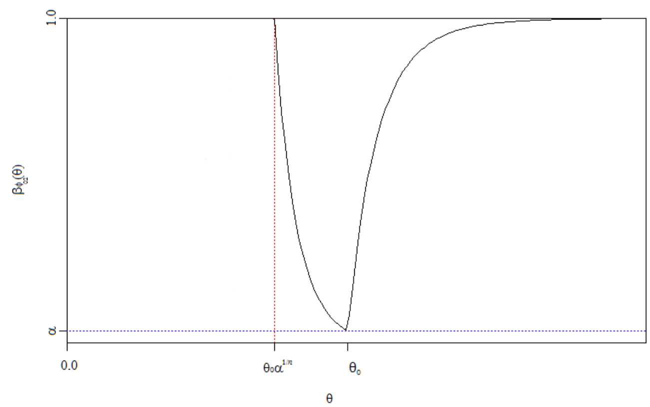
Comments on Test (11)

As we have seen, the test (11), though, UMP (at level α) for testing

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta > \theta_0,$$

is not UMP for testing

$$H_0' : \theta \leqslant \theta_0 \text{ against } H_1 : \theta > \theta_0.$$



However, as we shall see, later on, it is UMP for tesing $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$. As we will need this test later on, we assign an identifying name to it: ψ. Thus, from (11)

$$\psi(\mathbf{x}) = \begin{cases} 1, & \text{if } x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha} \\ 0, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < x_{(n)} \leqslant \theta_0 \\ 1, & \text{if } x_{(n)} > \theta_0. \end{cases} \tag{12}$$

Statistical Inference I

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

Hypothesis Testing in Uniform [0,θ] - III

Another Testing Problem

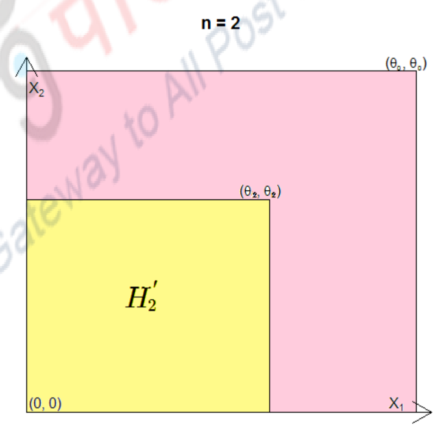
Under the same set-up, of $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ Unif $[0, \theta], \theta > 0$, we, now, carry out a test for

$$H_0 : \theta = \theta_0 \text{ (known) against } H_2 : \theta < \theta_0.$$

Start with

$$H_0 : \theta = \theta_0$$

versus $H_2' : \theta = \theta_2\,(< \theta_0)$. We aim to find a most powerful (MP) test for the problem of testing H_0 against H_2'. The following figure explains the hypotheses H_0 and H_2'. The region in yellow is the sample space under H_2'.



Mathematically, the pink (P) and yellow (Y) zones can be identified as

$$Y := \{\mathbf{x} \colon x_{(n)} \leqslant \theta_2\},$$

and

$$P := \{\mathbf{x} \colon \theta_2 < x_{(n)} \leqslant \theta_0\}.$$

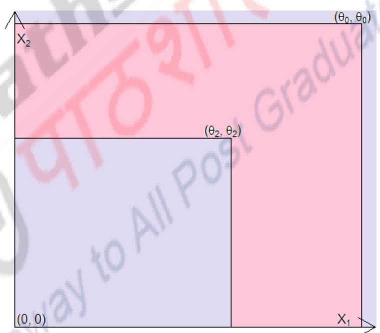
Derivation of MP Tests

$$p_{\theta_2}(\mathbf{x}) = \begin{cases} \frac{1}{\theta_2^n}, & \text{if } x_{(n)} \leqslant \theta_2 \\ 0, & \text{if } x_{(n)} > \theta_2. \end{cases} \tag{1}$$

$$k \cdot p_{\theta_0}(\mathbf{x}) = \begin{cases} \frac{k}{\theta_0^n}, & \text{if } x_{(n)} \leqslant \theta_0 \\ 0, & \text{if } x_{(n)} > \theta_0. \end{cases} \tag{2}$$

Take $k(\geqslant 0)$ such that

$$\frac{1}{\theta_2^n} = \frac{k}{\theta_0^n}$$
$$\Leftrightarrow k = \left(\frac{\theta_0}{\theta_2}\right)^n.$$

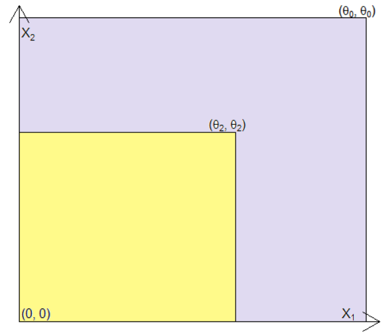


$p_{\theta_2}(\mathbf{x}) > k \cdot p_{\theta_0}(\mathbf{x}) \Leftrightarrow p_{\theta_2}(\mathbf{x}) = \frac{1}{\theta_2^n}$ and $k \cdot p_{\theta_0}(\mathbf{x}) = 0$, i.e., $x_{(n)} \leqslant \theta_2$ and $x_{(n)} > \theta_0$.

The regions in violet depict what the inequalities above say. Clearly, both inequalities cannot, simultaneously, hold. $p_{\theta_2}(\mathbf{x}) < k \cdot p_{\theta_0}(\mathbf{x}) \Leftrightarrow p_{\theta_2}(\mathbf{x}) = 0$ and $k \cdot p_{\theta_0}(\mathbf{x}) = \frac{k}{\theta_0^n}$, i.e., $\theta_2 < x_{(n)} \leqslant \theta_0$.

The region in violet depicts what the inequalities above say.

An MP Test:



By the Neyman-Pearson lemma, an MP at level α test for testing H_0 against $H_2' : \theta = \theta_2$, is of the form,

$$\phi^*(\mathbf{x}) = \begin{cases} 0, & \text{if } \theta_2 < x_{(n)} \leqslant \theta_0 \\ \text{any value in } [0,1], & \text{if } x_{(n)} > \theta_0 \text{ or } x_{(n)} \leqslant \theta_2, \end{cases} \tag{3}$$

such that

$$E_{\theta_0}\big[\phi^*(\mathbf{X})\big] = \alpha. \tag{4}$$

Non-randomized MP Tests:

To yield a non-randomized UMP test for testing H_0 against $H_2 : \theta < \theta_0$, we modify ϕ^* to (assume $\theta_2 \geqslant \theta_0 \cdot \sqrt[n]{\alpha}$)

$$\phi^*(\mathbf{x}) = \begin{cases} 0, & \text{if } \theta_2 < x_{(n)} \leqslant \theta_0 \\ 0, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < x_{(n)} \leqslant \theta_2 \\ 1, & \text{if } x_{(n)} > \theta_0 \text{ or } x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha}. \end{cases} \tag{5}$$

Combining the first two contingencies, we get

$$\phi^*(\mathbf{x}) = \begin{cases} 0, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < x_{(n)} \leqslant \theta_0 \\ 1, & \text{if } x_{(n)} > \theta_0 \text{ or } x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha}, \end{cases} \tag{6}$$

which is the same thing as ψ, as defined in (12).

Now, we show that ψ is indeed UMP for testing H_0 against $H_2 : \theta < \theta_0$. The test ψ (as given by (12))

$$\psi(\mathbf{x}) = \begin{cases} 0, & \text{if } \theta_0 \cdot \sqrt[n]{\alpha} < x_{(n)} \leqslant \theta_0 \\ 1, & \text{if } x_{(n)} > \theta_0 \text{ or } x_{(n)} \leqslant \theta_0 \cdot \sqrt[n]{\alpha} \end{cases} \tag{7}$$

is independent of θ_2, and gives a UMP test at level α for testing H_0 against $H_3 : \theta_0 \cdot \sqrt[n]{\alpha} \leqslant \theta < \theta_0$. For $\theta_2 < \theta_0 \cdot \sqrt[n]{\alpha}$, note that ψ has power 1 (maximum) for every $\theta < \theta_0 \cdot \sqrt[n]{\alpha}$, and is, hence, UMP for testing H_0 against $H_4 : \theta < \theta_0 \cdot \sqrt[n]{\alpha}$, as well. It is possible to identify a constant k, such that an application of the Neyman-Pearson lemma gives the UMP test for testing H_0 against H_4, without any reference to the power function. The question is can we guess what that constant is?

Thus, ψ is UMP at level α for testing $H_0 : \theta = \theta_0$ against both of $H_3 : \theta_0 \cdot \sqrt[n]{\alpha} \leqslant \theta < \theta_0$ and $H_4 : \theta < \theta_0 \cdot \sqrt[n]{\alpha}$. It is, therefore, UMP at level α for testing $H_0 : \theta = \theta_0$ against $H_2 : \theta < \theta_0$.

UMP test for H_0 versus $H_A : \theta \neq \theta_0$:

As it is, also, UMP at level α for testing

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta > \theta_0,$$

we conclude that ψ is UMP at level α for testing

$$H_0 : \theta = \theta_0 \text{ against } H_A : \theta \neq \theta_0.$$

Final Comments:

In the case of $\mathsf{Unif}\,[0, \theta]$, for the problem of testing

$$H_0 : \theta = \theta_0 \text{ against } H_a : \theta \neq \theta_0$$

at level α, we have obtained a UMP test. This, however, is a rare instance of the situation where a UMP test for (H_0, H_a) exists. Another situation where a UMP test for (H_0, H_a) exists, is for the exponential $\mathsf{Exp}(\theta, 1)$ distribution, with p.d.f.,

$$f_\theta(x) = \begin{cases} e^{-(x-\theta)}, & \text{if } x > \theta \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

If, somehow, the $\mathsf{Exp}\,(\theta, 1)$ distribution can be transformed to $\mathsf{Unif}\,[0, \theta']$, our problem would reduce to the case discussed. Can you find some function g, such that for $X \sim \mathsf{Exp}\,(\theta, 1)$, $g(X) \sim \mathsf{Unif}[0, \theta']$?

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Hypothesis testing in Shifted Exponential Population
Module- 29

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In the last three modules we have discussed the UMP test relating to $R(0,\theta)$ population. We have considered both one-sided and two sided alternatives. In the present module we are going to discuss the the UMP test for shifted exponential distributions. We consider both the one parameter and two parameter cases. At first we discuss the single parameter case where we use the UMP test derived for the $R(0,\theta)$ population.

One-parameter exponential distribution

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from the population with p.d.f.

$$\begin{aligned} f(x,\mu) &= \exp^{-(x-\mu)}, x \geq \mu \\ &= 0, \text{otherwise}, \end{aligned}$$

with $-\infty < \mu < \infty$. Also the joint density for $X_1, X_2, \ldots, X_n$ is given as,

$$\begin{aligned} p(\mathbf{x},\mu) &= \exp^{-\sum_{i=1}^{n}(x_i-\mu)}, x_i \geq \mu, \forall\ i \\ &= 0, \text{otherwise}. \end{aligned}$$

If we define $x_{(1)}$ as the smallest order statistics and notice that $x_i > \mu, \forall\ i \Leftrightarrow\ x_{(1)} > \mu$. Hence the above joint density can be rewritten as,

$$\begin{aligned} p(\mathbf{x},\mu) &= \exp^{-\sum_{i=1}^{n}(x_i-\mu)}, x_{(i)} \geq \mu \\ &= 0, \text{otherwise}. \end{aligned}$$

Suppose our object is to find the UMP test for

$$H_0 : \mu = \mu_0 \text{ vs. } H_A : \mu \neq \mu_0$$

. Consider the transformation

$$Y_i = e^{-X_i}, i = 1, 2, \ldots n.$$

Then $Y_i, i = 1, 2, \ldots n$ are independently distributed $\sim R(0, \theta)$ random variables where $\theta = e^{-\mu}$. Based on the observations $Y_i, i = 1, 2, \ldots n$, UMP size α test for $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$ is given by,

$$\begin{aligned} \varphi(\mathbf{Y}) &= 0 \text{ if } e^{-\mu_0}\alpha^{\frac{1}{n}} \leq Y_{(n)} \leq e^{-\mu_0} \\ &= 1 \text{ otherwise.} \end{aligned}$$

In terms of the observations $X_i, i = 1, 2, \ldots n$ the UMP test is given by

$$\begin{aligned} \varphi(\mathbf{X}) &= 0 \text{ if } \mu_0 \leq X_{(1)} \leq \mu_0 - \frac{1}{n}\log\alpha \\ &= 1 \text{ otherwise.} \end{aligned}$$

Two-parameter exponential distribution

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from the population with p.d.f.

$$\begin{aligned} f(x, \mu, \sigma) &= \frac{1}{\sigma}\exp^{\frac{-(x-\mu)}{\sigma}}, x \geq \mu \\ &= 0, \text{otherwise,} \end{aligned}$$

with $-\infty < \mu < \infty$ and $\sigma > 0$. Defining $\theta = (\mu.\sigma)$ the joint density is given as,

$$\begin{aligned} p(\mathbf{x}, \theta) &= \frac{1}{\sigma^n}\exp^{-\frac{1}{\sigma}\sum_{i=1}^{n}(x_i-\mu)}, x_i \geq \mu, \forall\, i \\ &= 0, \text{otherwise,} \end{aligned}$$

with $-\infty < \mu < \infty$ and $\sigma > 0$. Therefore the parametric space is $\Theta = (-\infty, \infty) \times (0, \infty)$. Consider the following testing problems,

$$I. H_0 : \mu = \mu_0, \sigma = \sigma_0 \text{ vs. } H_A : \mu < \mu_0, \sigma < \sigma_0$$

$$II. H_0 : \mu = \mu_0, \sigma = \sigma_0 \text{ vs. } H_A : \mu > \mu_0, \sigma < \sigma_0$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Solution.(I) Since $x_i > \mu$,$\forall\ i \Leftrightarrow x_{(1)} > \mu$, $f(\mathbf{x}, \theta) = 0, \Leftrightarrow\ x_{(1)} < \mu$. Taking $(\mu_1, \sigma_1) : \mu_1 < \mu_0, \sigma_1 < \sigma_0$, we have

$$\begin{aligned} p(\mathbf{x}, \theta_1) &\geq 0, x_{(1)} > \mu_1 \\ &= 0, x_{(1)} > \mu_1 \end{aligned}$$

$$\begin{aligned} p(\mathbf{x}, \theta_0) &\geq 0, x_{(1)} > \mu_0 \\ &= 0, x_{(1)} > \mu_0. \end{aligned}$$

Now whenever $x_{(1)} > \mu_0$ we have,

$$p(x, \theta_1), p(x, \theta_0) > 0.$$

Thus we get

$$\begin{aligned} \frac{p(x, \theta_1)}{p(x, \theta_0)} &= \text{finite}, \ \text{ if } x_{(1)} > \mu_0 \\ &= \infty, \ \text{if} x_{(1)} < \mu_0 \end{aligned}$$

Hence, for testing $H_A : \mu = \mu_1 (< \mu_0), \sigma = \sigma_1 (< \sigma_0)$ the MP size α test has a critical region,

$$\omega = \left\{ x : x_{(1)} < \mu_0 \text{ or } x_{(1)} > \mu_0 \ \text{ and } \ \frac{p(x, \mu_1, \sigma_1)}{p(x, \mu_0, \sigma_0)} > k \right\},$$

where $k (\geq 0)$ such that, $P_{H_0}(\omega) = \alpha$. Now,

$$\begin{aligned} &\frac{p(x, \mu_1, \sigma_1)}{p(x, \mu_0, \sigma_0)} \\ &= \left(\frac{\sigma_0}{\sigma_1} \right)^n exp \left\{ \left(\frac{1}{\sigma_0} - \frac{1}{\sigma_1} \right) \sum_{i=1}^{n} x_i + n \left(\frac{\mu_1}{\sigma_1} - \frac{\mu_0}{\sigma_0} \right) \right\} \end{aligned}$$

where $x_{(1)} \geq \mu_0$. The above ratio is non-increasing in $\sum_{i=1}^{n} x_i$ so long as $\sigma_1 < \sigma_0$. Hence there is a constant c such that,

$$\frac{p(x, \theta_1)}{p(x, \theta_0)} > k \Leftrightarrow \sum_{i=1}^{n} x_i < c.$$

Thus ω is equivalent to,

$$\omega = \left\{ x : x_{(1)} < \mu_0 \text{ or } x_{(1)} \geq \mu_0 \text{ and } \sum_{i=1}^{n} x_i < c \right\},$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

where c is such that,

$$P_{\theta_0}\left\{X_{(1)} < \mu_0\right) + P_{\theta_0}\left\{X_{(1)} \geq \mu_0 \text{ and } \sum_{i=1}^{n} X_i < c)\right\} = \alpha.$$

Since $P_{\theta_0}\left\{X_{(1)} < \mu_0\right) = 0$, we get

$$\begin{aligned} \int_{x_{(1)}\geq\mu_0, \sum_{i=1}^{n} x_i < c} p(\mathbf{x}, \mu_0, \sigma_0) d\mathbf{x} &= \alpha \\ i.e., \int_{\sum_{i=1}^{n} x_i < c} p(\mathbf{x}, \mu_0, \sigma_0) d\mathbf{x} &= \alpha. \end{aligned}$$

If we consider the transformation, $Y = \frac{2}{\sigma_0}\sum_{i=1}^{n}(x_i - \mu)$, therefore $Y \sim \chi^2_{2n}$. Hence,

$$\int_0^{2\frac{(c-n\mu_0)}{\sigma_0}} f_{\chi^2_{2n}}(y)dy = \alpha,$$

and if we denote the lower α point of χ^2_{2n} is given by $\chi^2_{2n,\alpha}$, then, $c = \frac{\sigma_0\chi^2_{2n,\alpha}}{2} + n\mu_0$. Therefore the MP test is given by,

$$\omega = \left\{x : x_{(1)} < \mu_0 \text{ or } x_{(1)} \geq \mu_0 \text{ and, } \sum_{i=1}^{n} x_i < \frac{\sigma_0\chi^2_{2n,\alpha}}{2} + n\mu_0\right\},$$

which is independent of any $(\mu_1, \sigma_1); \mu_1 < \mu_0, \sigma_1 < \sigma_0$. Hence it is an UMP test as well.

Solution.(II) We take $(\mu_1, \sigma_1) : \mu_1 > \mu_0, \sigma_1 < \sigma_0$, Thus we get

$$\begin{aligned} \frac{p(x,\theta_1)}{p(x,\theta_0)} &= \text{finite, if } x_{(1)} < \mu_1 \\ &= 0, \text{ if} x_{(1)} > \mu_1 \end{aligned}$$

Hence, for testing $H_A : \mu = \mu_1(> \mu_0), \sigma = \sigma_1(< \sigma_0)$ the MP size α test has a critical region,

$$\omega = \left\{x : x_{(1)} > \mu_1 \text{ and } \frac{p(x,\theta_1)}{p(x,\theta_0)} > k\right\},$$

where $k(\geq 0)$ such that, $P_{H_0}(\omega) = \alpha$. Now as in case (I),

$$\frac{p(x,\theta_1)}{p(x,\theta_0)} > k \Leftrightarrow \sum_{i=1}^{n} x_i < c.$$

Thus ω is equivalent to,

$$\omega = \left\{ x : x_{(1)} > \mu_1 \text{ and } \sum_{i=1}^{n} x_i < c \right\},$$

where c is such that,

$$P_{\theta_0} \left\{ X_{(1)} > \mu_1 \right) + P_{\theta_0} \left\{ \sum_{i=1}^{n} X_i < c) \right\} = \alpha,$$

$$\Rightarrow e^{\frac{-n(\mu_0 - \mu_1)}{\sigma_0}} + P_{\theta_0} \left[\chi^2_{2n} < \frac{2(c - n\mu_0)}{\sigma_0} \right] = \alpha,$$

$$\Rightarrow c = n\mu_0 + \frac{\sigma_0}{2} \left[\alpha - e^{\frac{-n(\mu_0 - \mu_1)}{\sigma_0}} \right] \chi^2_{1-\alpha;2n}.$$

Since the non-randomized MP test depends upon a specific alternative value μ_1, it is not UMP for testing $H_0 : \mu = \mu_0, \sigma = \sigma_0$ vs. $H_A : \mu > \mu_0, \sigma < \sigma_0$. Hence no UMP test exists for this problem.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Testing of Composite Null Hypotheses against Simple Alternatives
Module- 30

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

$\{p(x,\theta),\theta\in\Theta\}$, $\Theta\in R^p, p\geq 2$. Testing problem :

$$H:\theta\in\Theta_H \text{ vs. } H_A:\theta=\theta_1 \text{ (known) } \notin\Theta_H$$

Define,

$$\begin{aligned} M(\alpha,H) &: \text{ class of all level } \alpha \text{ for testing H} \\ &\equiv \{\varphi: E_\theta\varphi(X)\leq\alpha,\theta\in\Theta_H\}\end{aligned}$$

Let $\omega(\theta)$ be a distribution of θ over Θ_H. $\omega(\theta)$ is known. $\omega(\theta)\geq 0\ \forall\theta\in\Theta_H$.

$$\int_{\Theta_H}\omega(\theta)d\mu(\theta)=1 \;\; [\mu(\theta): \text{Measure on}\Theta_H]$$

$$p(x,\theta): \text{p.d.f./p.m.f. of} X \text{ given } \theta$$

$$p(x,\theta)\omega(\theta): \text{ Joint p.d.f. / p.m.f. of } (X,\theta) \text{ over } \chi\times\Theta_H.$$

Marginal distribution of X is,

$$\begin{aligned} g_\omega(x) &= \int_{\Theta_H} p(x,\theta)\omega(\theta)d\mu(\theta), x\in\chi \\ &= \text{known, when } \omega(\theta) \text{ is known apriori.}\end{aligned}$$

Next consider the testing problem,

$$H_\omega: X\sim g_\omega(x) \text{ vs. } H_A:\theta=\theta_1.$$

Hence, H_ω and H_A are both simple. Therefore, NP lemma, MP test for H_ω vs. H_A is defined by the function,

$$\begin{aligned}
\varphi_\omega(x) &= 1 \text{ if } p(x,\theta_1) > kg_\omega(x) \\
&= a \text{ if } p(x,\theta_1) = kg_\omega(x) \\
&= 0 \text{ if } p(x,\theta_1) < kg_\omega(x).
\end{aligned}$$

with $E_{H_\omega}\varphi_\omega(X) = \alpha$. Suppose such test satisfies,

$$E_\theta\varphi_\omega(X) \leq \alpha \ \ \forall\theta \in \Theta_H.$$

Then, such φ_ω is MP size α for (H, H_A),

$$\begin{aligned}
M(\alpha, H_\omega) &: \text{ class of all level } \alpha \text{ for testing } H_\omega \\
&\equiv \{\varphi : E_{H_\omega}\varphi(X) \leq \alpha\}.
\end{aligned}$$

It can be shown that,

$$M(\alpha, H) \subset M(\alpha, H_\omega)$$

Proof. $\varphi \in M(\alpha, H)$

$$\begin{aligned}
&\Rightarrow E_\theta\varphi(X) \leq \alpha \\
&\Leftrightarrow \int_\chi \varphi(x)p(x,\theta)d\mu(x) \leq \alpha \ \ \forall\theta \in \Theta_H \\
&\Rightarrow \int_{\Theta_H}\int_\chi \varphi(x)p(x,\theta)\omega(\theta)d\mu(x)d\mu(\theta) \leq \alpha \\
&\Leftrightarrow \int_\chi \varphi(x)\left[\int_{\Theta_H} p(x,\theta)\omega(\theta)d\mu(\theta)\right]d\mu(x) \leq \alpha \text{ (By Fubini's Theorem)} \\
&\Leftrightarrow \int_\chi \varphi(x)g_\omega(x)d\mu(x) \leq \alpha \\
&\Leftrightarrow E_{H_\omega}\varphi(X) \leq \alpha \\
&\Leftrightarrow \varphi \in M(\alpha, H_\omega).
\end{aligned}$$

Conclusions:

$$(i)\varphi_\omega : \text{ MP size } \alpha \text{ for } (H_\omega, H_A)$$

$$(ii)E_\theta\varphi_\omega(X) \leq \alpha \ \forall\theta \in \Theta_H$$

$$(iii)M(\alpha, H) \subset M(\alpha, H_\omega)$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

To conclude that, φ_ω is MP size α for (H, H_A). We can use the following argument, φ_ω is MP within $M(\alpha, H_\omega)$ (by (i))

$$\Rightarrow E_{\theta_1}\varphi_\omega(X) \geq E_{\theta_1}\varphi(X) \forall \varphi : \varphi \in M(\alpha, H_\omega)$$
$$\Rightarrow E_{\theta_1}\varphi_\omega(X) \geq E_{\theta_1}\varphi(X) \forall \varphi : \varphi \in M(\alpha, H) (\text{ by } (iii))$$

Again, by (ii), $\varphi_\omega \in M(\alpha, H)$, we have, φ_ω is MP for (H, H_A) within $M(\alpha, H)$.
Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ independently $\theta = (\mu, \sigma) \in (R, R^+) \subset R^2$. Testing problem,

$$H_1 : \left[\sigma^2 \geq \sigma_0^2, -\infty < \mu < \infty\right] \text{ vs. } H_{A_1} : \left[\sigma^2 = \sigma_1^2 (< \sigma_0^2), \mu = \mu_1\right]$$

$$H_2 : \left[\sigma^2 \leq \sigma_0^2, -\infty < \mu < \infty\right] \text{ vs. } H_{A_1} : \left[\sigma^2 = \sigma_1^2 (> \sigma_0^2), \mu = \mu_1\right]$$

Solution: Sufficient statistic for (μ, σ^2) is, $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n (X_i - \bar{X})^2\right) = (u, v)$. So, test based on $\mathbf{X} = (X_1, \ldots, X_n)$ has the same power as that based on (u, v). For each test $\varphi = \varphi(X_1, \ldots, X_n)$, we can find another test $\psi = \psi(u, v) = \psi(T)$ such that φ and ψ have identical power functions. So, WLOG, we consider to be dependent on T, if required. Let $\omega(\theta)$ be a non-negative weight function on $\sigma^2 = \sigma_0^2$, $i.e.$, $\omega(\theta) = \omega(\mu, \sigma_0^2), -\infty < \mu < \infty$ with the propoerty that,

$$(i) \omega(\mu, \sigma_0^2) \geq 0 \forall \mu$$

$$(ii) \int_{-\infty}^{\infty} \omega(\mu, \sigma_0^2) d\mu = 1$$

Here we have the following likelihood functions,

$$\begin{aligned} p(x, \theta) &= c\sigma^{-n} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right] \\ &= c\sigma^{-n} exp\left[-\frac{v}{2\sigma^2}\right] exp\left[-\frac{n}{2\sigma^2}(u - \mu)^2\right] \end{aligned}$$

We know U and V are independently distributed, where $U \sim N(\mu, \frac{\sigma^2}{n})$ and $\frac{V}{\sigma^2} \sim \chi^2_{n-1}$.

$$p(u, v, \theta) = c_1 \sigma^{-n} v^{\frac{n-3}{2}} exp\left[-\frac{v}{2\sigma^2}\right] exp\left[-\frac{n}{2\sigma^2}(u - \mu)^2\right].$$

Since, $H_\omega : X \sim g_\omega(x)$, where $g_\omega(x) = \int_{-\infty}^{\infty} p(x,\theta)\omega(\mu,\sigma_0^2)d\mu$, therefore,

$$\begin{aligned} p(u,v,H_\omega) &= c_1\sigma^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma^2}\right]\int_{-\infty}^{\infty} exp\left[-\frac{n}{2\sigma^2}(u-\mu)^2\right]\omega(\mu,\sigma_0^2)d\mu \\ &= \text{known.} \end{aligned} \tag{0.1}$$

Therefore,

$$p(x,\theta_1) = p(u,v,\theta_1) = c_1\sigma_1^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma_1^2}\right]exp\left[-\frac{n}{2\sigma_1^2}(u-\mu_1)^2\right].$$

Hence the likelihood ratio is given by,

$$R = \frac{c_1\sigma_1^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma_1^2}\right]exp\left[-\frac{n}{2\sigma_1^2}(u-\mu_1)^2\right]}{c_1\sigma^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma^2}\right]\int_{-\infty}^{\infty} exp\left[-\frac{n}{2\sigma^2}(u-\mu)^2\right]\omega(\mu,\sigma_0^2)d\mu}.$$

Consider the following weight function,

$$\begin{aligned} \omega(\mu,\sigma^2) &= 1 \text{ if } (\mu,\sigma^2) = (\mu,\sigma_0^2) \\ &= 0, \text{ otherwise.} \end{aligned}$$

Then R boils down to,

$$\begin{aligned} R &= \frac{c_1\sigma_1^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma_1^2}\right]exp\left[-\frac{n}{2\sigma_1^2}(u-\mu_1)^2\right]}{c_1\sigma_0^{-n}v^{\frac{n-3}{2}}exp\left[-\frac{v}{2\sigma_0^2}\right]exp\left[-\frac{n}{2\sigma_0^2}(u-\mu_1)^2\right]} \\ &= c^\star exp\left\{\left(-\frac{1}{2\sigma_1^2}+\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right)\right\}. \end{aligned}$$

which increases or decreases with $\sum_{i=1}^{n}(x_i-\mu_1)^2$ according as $\sigma_1^2 >$ or $< \sigma_0^2$. Hence using the weight function, a MP test for H_ω against H_A has the critical region,

$$\omega_1 = \left\{x : \sum_{i=1}^{n}(x_i-\mu_1)^2 > k_1\right\}$$

or,

$$\omega_2 = \left\{x : \sum_{i=1}^{n}(x_i-\mu_1)^2 < k_2\right\}$$

according as $\sigma^2 >$ or $< \sigma_0^2$, where k_1(or k_2) is such that $P_{H_\omega}(\omega_1) = P_{H_\omega}(\omega_2) = \alpha$. So,

$$\text{Under } H_\omega,\ \ \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{\sigma_0^2} \sim \chi_n^2$$

and k_1 and k_2 can be evaluated from the percentile points of χ^2_n distribution. Now, we have to veify that, under the above weight function,

$$P_\theta(\omega_k) \leq \alpha, \;\; \forall\theta \in \Theta_H, k = 1, 2$$

Here,

$$\begin{aligned} P_\theta(\omega_1|(\mu,\sigma^2)) &= P_\theta\left\{\frac{\sum_{i=1} n(X_i-\mu_1)^2}{\sigma^2} \geq \frac{k_1}{\sigma^2}|(\mu,\sigma^2)\right\} \\ &= P_\theta\left\{Y \geq \frac{k_1}{\sigma^2}|(\mu,\sigma^2)\right\} \end{aligned}$$

where $Y \sim$ non-central χ^2 with $d.f. = n$ and n.c.p. $\Delta^2 = \frac{n(\mu_1-\mu)^2}{\sigma^2}$. It can be seen that,

$$P(Y > c) \uparrow \Delta^2 \text{ (for fixed } c \text{ and } n)$$

which implies,

$$P\left[Y > c|\Delta^2 > 0\right] \geq P\left[Y > c|\Delta^2 = 0\right]$$

Hence we get,

$$P_\theta(\omega|\theta) \leq \alpha, \forall\theta \in \Theta_H,$$

and we conclude, the test ω_1 is not MP size α for

$$H : \left[-\infty < \mu < \infty, \sigma^2 \leq \sigma_0^2\right] \text{ vs. } H_A : \left[\mu = \mu_1, \sigma^2 = \sigma_1^2(> \sigma_0^2)\right].$$

Let us now consider,

$$\begin{aligned} P\{\omega_2|(\mu,\sigma^2)\} = P\{Y < \frac{k_2}{\sigma^2}|(\mu,\sigma^2)\} &< P\{Y < \frac{k_2}{\sigma^2}|\Delta^2 = 0\} \\ &< P\{Y < \frac{k_2}{\sigma_0^2}|\Delta^2 = 0\}\forall\sigma^2 > \sigma_0^2 \\ &= \alpha. \end{aligned}$$

Thus, ω_2 is MP for,

$$H : \left[-\infty < \mu < \infty, \sigma^2 \geq \sigma_0^2\right] \text{ vs. } H_A : \left[\mu = \mu_1, \sigma^2 = \sigma_1^2(< \sigma_0^2)\right].$$

Alternatively, set

$$\begin{aligned}\omega^\star(\theta)=\omega^\star(\mu,\sigma^2) &= \frac{1}{\delta\sqrt{2\pi}}exp\left\{-\frac{1}{2\delta^2}(\mu-\mu_1)^2\right\}, -\infty<\mu<\infty \text{ at } \sigma^2=\sigma_0^2\\ &= 0 \text{ otherwise.}\end{aligned}$$

$$\bar{X}|(\mu,\sigma_0^2)\sim N\left(\mu,\frac{\sigma_0^2}{n}\right)$$

$$\mu\sim N(\mu,\delta^2)$$

$$\bar{X}\sim N\left(\mu_1,\delta^2+\frac{\sigma_0^2}{n}\right)$$

To choose δ^2 such that, $\delta^2+\frac{\sigma_0^2}{n}=\frac{\sigma_1^2}{n}\Rightarrow\delta^2=\frac{1}{n}(\sigma_1^2-\sigma_0^2)$. Thus,

$$\bar{X}|\omega^\star\sim N\left(\mu,\frac{\sigma_1^2}{n}\right)$$

Here, we find MP size α test for $H_\omega:[X\sim p_{\omega^\star}(\theta)=c_1\times exp\left\{-\frac{v}{2\sigma_0^2}\right\}exp\left\{-\frac{n}{2\sigma_1^2}(u-\mu_1)^2v^{\frac{n-3}{2}}\right\}]$ vs. $H_A:[\mu=\mu_1,\sigma^2=\sigma_1^2>\sigma_0^2]$. The likelihood ratio for the above test is given by,

$$\frac{p(x|H_A)}{p(x|H_{\omega^\star})}=\frac{c_1\times exp\left\{-\frac{v}{2\sigma_1^2}\right\}exp\left\{-\frac{n}{2\sigma_1^2}(u-\mu_1)^2v^{\frac{n-3}{2}}\right\}}{c_1\times exp\left\{-\frac{v}{2\sigma_0^2}\right\}exp\left\{-\frac{n}{2\sigma_1^2}(u-\mu_1)^2v^{\frac{n-3}{2}}\right\}}=c\times exp\left\{-\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_0^2}\right)v\right\}\uparrow v.$$

Hence, MP test for $(H_{\omega^\star},H_A)$ has the critical region,

$$\omega_3=\{x;\sum_{i=1}^n(x_i-\bar{x})^2>k_3\},$$

where, k_3 is such that, $P_{H_{\omega^\star}}=\alpha\Rightarrow k_3=\chi^2_{n-1,\alpha}\sigma_0^2$. Also it can be shown that,

$$P\{\omega_3|\theta\}\le\alpha,\forall\theta=(\mu,\sigma^2);-\infty<\mu<\infty,\sigma^2\le\sigma_0^2.$$

Now,

$$\begin{aligned}P\left[\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{\sigma^2}>\frac{k_3}{\sigma^2}|(\mu,\sigma^2)\right] &= P\left[\chi^2_{n-1}>\chi^2_{n-1,\alpha}\frac{\sigma_0^2}{\sigma^2}\right]\quad\forall-\infty<\mu<\infty\\ &\le P\left[\chi^2_{n-1}>\chi^2_{n-1,\alpha}\right]\quad\forall-\infty<\mu<\infty,\sigma^2<\sigma_0^2\\ &= \alpha.\end{aligned}$$

The above test is MP size α for

$$H:\left[-\infty<\mu<\infty,\sigma^2\le\sigma_0^2\right]\text{ vs. }H_A:\left[\mu=\mu_1,\sigma^2=\sigma_1^2(>\sigma_0^2)\right].$$

Since ω_3 is independent of H_A, the test is also UMP size α for,

$$H : \left[-\infty < \mu < \infty, \sigma^2 \le \sigma_0^2\right] \text{ vs. } H_A : \left[-\infty < \mu < \infty, \sigma^2 = \sigma_1^2 (> \sigma_0^2)\right].$$

ePathshala पाठशाला
A Gateway to All Post Graduate Courses

Statistical Inference I
Monotone Likelihood Ratio

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In the present module we define the monotone likelihood ratio (MLR) property for a family of pmf or pdf denoted by $\{p(x, \theta) : \theta \in \Theta \subset \mathcal{R}\}$. we exploit this property to derive the UMP level α tests for one-sided null against onesided alternative hypotheses in some situations.

Monotone Likelihood Ration (MLR) family of distribution

A real parametric family $\{p(x, \theta) : \theta \in \Theta \subset \mathcal{R}\}$ is said to have MLR property in a real valued statistic $T(x)$ if, for any $\theta_1 < \theta_2 \in \Theta$, the following are satisfied.

$i) p(x, \theta_1) \neq p(x, \theta_2)$

[Distribution are distinct corresponding to distinct parameter points]

$ii)$ The ratio $R(x) = \dfrac{p(x, \theta_2)}{p(x, \theta_1)}$

is non-decreasing in $T(x)$ on the set $\{x : max(p(x, \theta_2), p(x, \theta_1)) > 0\}$.

Note If $p(x, \theta_2) = 0$ and $p(x, \theta_1) > 0$, $R(x) = 0$.

$p(x, \theta_2) > 0$ and $p(x, \theta_1) = 0$, $R(x) = \infty$.

Some examples on MLR families

1. $\{p(x, \theta), \theta \in \Theta \subset R\}$: One parameter Exponential family. Then we can express $p(x, \theta)$ in the form,

$$p(x, \theta) = u(\theta) exp(q(\theta)T(x))v(x)$$

such that $u(\theta)$ and $q(\theta)$ depends only on θ, $v(x)$ is independent of θ and $T(x)$ depends only on x. We set $T(x)$ such that $Q(\theta)$ is a strictly increasing function of θ. Then we have for $\theta_1 < \theta_2$,

$$\frac{p(x, \theta_2)}{p(x, \theta_1)} = \frac{u(\theta_2)}{u(\theta_1)} exp\{(Q(\theta_2) - Q(\theta_1))T(x)\},$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

increasing in $T(x)$ because $Q(\theta)$ is a strictly increasing function of θ . Hence, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x)$.

Note If $(X_1, X_2, \ldots, X_n)$ is a random sample of size n from the population with p.m.f or p.d.f. $p(x,\theta)$ then $p(x,\theta)$ has MLR in $\sum_{i=1}^{n} T(x_i)$.

Consider the following examples.

1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $N(\theta, 1)$, population. Therefore,

$$\begin{aligned}
p(\mathbf{x},\theta) &= (2\pi)^{-n/2} exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\\
&= exp\left\{-\frac{n}{2}\theta^2\right\} exp\left\{\theta\sum_{i=1}^{n}x_i\right\}(2\pi)^{-n/2}exp\left\{-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right\}\\
&= u(\theta)exp(q(\theta)T(\mathbf{x}))v(\mathbf{x})
\end{aligned}$$

where $u(\theta) = exp\left\{-\frac{n}{2}\theta^2\right\}$, $Q(\theta) = \theta$, $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$
and $v(\mathbf{x}) = (2\pi)^{-n/2} exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}$.
$p(\mathbf{x},\theta)$ has MLR in $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$.

2. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample of size n from $N(0, \theta^2)$, population. Therefore,

$$\begin{aligned}
p(\mathbf{x},\theta) &= (2\pi)^{-n/2}\theta^{-n} exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)\\
&= u(\theta)exp(q(\theta)T(\mathbf{x}))v(\mathbf{x})
\end{aligned}$$

where $u(\theta) = \theta^{-n}$, $q(\theta) = -\frac{1}{2\theta^2}$, $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and $v(\mathbf{x}) = (2\pi)^{-n/2}$.
$p(\mathbf{x},\theta)$ has MLR in $T(\mathbf{x}) = \sum_{i=1}^{n} x_i^2$.

3. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample of size n from $Bernoulli(\theta)$ population.

$$\begin{aligned}
p(\mathbf{x},\theta) &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\\
&= (1-\theta)^n exp\left[ln(\frac{\theta}{1-\theta})\sum_{i=1}^{n}x_i\right]\\
&= u(\theta)exp(q(\theta)T(\mathbf{x})v(\mathbf{x})
\end{aligned}$$

where $c(\theta) = \theta^n, q(\theta) = ln(\frac{\theta}{1-\theta}), T(\mathbf{x}) = \sum_{i=1}^n x_i$ and $v(\mathbf{x}) = 1$.

$p(\mathbf{x}, \theta)$ has MLR in $T(\mathbf{x}) = \sum_{i=1}^n x_i$.

4. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample of size n from the geometric distribution with p.m.f.

$$p(x, \theta) = \theta(1-\theta)^x, x = 0, 1, 2, \ldots, 0 < \theta < 1.$$

Then

$$\begin{aligned} p(\mathbf{x}, \theta) &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \theta^n exp\left[ln(1-\theta)\sum_{i=1}^n x_i\right] \\ &= u(\theta)exp(q(\theta)T(\mathbf{x})v(\mathbf{x}) \end{aligned}$$

where $c(\theta) = \theta^n, q(\theta) = -ln(1-\theta), T(\mathbf{x}) = -\sum_{i=1}^n x_i$ and $v(\mathbf{x}) = 1$.

$p(\mathbf{x}, \theta)$ has MLR in $T(\mathbf{x}) = -\sum_{i=1}^n x_i$.

5. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample of size n from the exponential distribution with p.d.f.

$$p(x, \theta) = \theta exp[-\theta x], x > 0, \theta > 0$$

Now

$$\begin{aligned} p(\mathbf{x}, \theta) &= \theta^n exp\left[-\theta\sum_{i=1}^n x_i\right] \\ &= u(\theta)exp(q(\theta)T(\mathbf{x})v(\mathbf{x}) \end{aligned}$$

where $u(\theta) = \theta^n, q(\theta) = \theta, T(\mathbf{x}) = -\sum_{i=1}^n x_i$ and $v(\mathbf{x}) = 1$.

$p(\mathbf{x}, \theta)$ has MLR in $T(\mathbf{x}) = -\sum_{i=1}^n x_i$.

6. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample of size n from the exponential distribution with p.d.f.

$$p(x, \theta) = \frac{1}{\theta}exp[-\frac{x}{\theta}], x > 0, \theta > 0$$

Now

$$\begin{aligned} p(\mathbf{x}, \theta) &= \left(\frac{1}{\theta}\right)^n exp\left[-\frac{\sum_{i=1}^n x_i}{\theta}\right] \\ &= u(\theta) exp(q(\theta) T(\mathbf{x}) v(\mathbf{x}) \end{aligned}$$

where $u(\theta) = \left(\frac{1}{\theta}\right)^n, q(\theta) = -\frac{1}{\theta}, T(\mathbf{x}) = \sum_{i=1}^n x_i$ and $v(\mathbf{x}) = 1$.

$p(\mathbf{x}, \theta)$ has MLR in $T(\mathbf{x}) = \sum_{i=1}^n x_i$.

Non-exponential family

1. $X \sim Cauchy(\theta, 1)$.

$$p(x, \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

For any $\theta_2 > \theta_1$

$$\begin{aligned} \frac{p(x, \theta_2)}{p(x, \theta_1)} = \frac{1 + (x - \theta_1)^2}{1 + (x - \theta_2)^2} &> 1 \quad \text{as } x \longrightarrow \theta_2 \\ &< 1 \quad \text{as } x \longrightarrow \theta_1 \\ &\longrightarrow 1 \text{ as } x \longrightarrow \pm\infty. \end{aligned}$$

Thus $Cauchy(\theta, 1)$ is not a member of MLR family.

2. $X \sim Cauchy(0, \theta)$

$$p(x, \theta) = \frac{1}{\pi} \frac{\theta}{\theta^2 + x^2}$$

for any $\theta_1 < \theta_2$,

$$\frac{p(x, \theta_2)}{p(x, \theta_1)} = \left(\frac{\theta_2^2}{\theta_1^2}\right) \frac{\theta_1^2 + x^2}{\theta_2^2 + x^2} = \left(\frac{\theta_2^2}{\theta_1^2}\right) \left[1 - \frac{\theta_2^2 - \theta_1^2}{\theta_2^2 + x^2}\right]$$

increasing in x^2 or in $|x|$, Thus $Cauchy(0, \theta)$ is a member of MLR family in $|x|$.

3. $X_i \sim R(0, \theta)$ for $i = 1, \ldots, n$ independently and let $\mathbf{X} = (X_1, \ldots, X_n)$

$$\begin{aligned} p(\mathbf{x}, \theta) &= \theta^{-n} \text{ if } 0 < x_{(n)} < \theta \\ &= 0 \text{ otherwise.} \end{aligned}$$

where $x_{(n)}$ is the largest order statistic. Take $\theta_1 < \theta_2$. Thus,

$$\begin{aligned}
\frac{p(x,\theta_2)}{p(x,\theta_1)} &= \left(\frac{\theta_2}{\theta_1}\right)^n \text{ if } 0 < x_{(n)} < \theta_1 \\
&= \infty \text{ if } \theta_1 < x_{(n)} < \theta_2.
\end{aligned}$$

is non-decreasing in $x_{(n)}$. Thus $R(0,\theta)$ is a member of MLR family in $x_{(n)}$.

4. $X \sim$ Hypergeometric (θ, n, N). θ : unknown, We know that,

$$p(x,\theta) = \frac{\binom{\theta}{x}\binom{N-\theta}{n-x}}{\binom{N}{n}}, \quad \max\{n-N+\theta, 0\} \leq x \leq \min\{\theta, n\}$$

and hence,

$$\begin{aligned}
\frac{p(x,\theta+1)}{p(x,\theta)} &= \frac{\theta+1}{\theta-x+1}\frac{N-n-\theta+x}{N-\theta} \\
&= \frac{\theta+1}{\theta-x+1}\frac{N-n-\theta+x}{N-\theta} \text{ if, } \theta+n-N+1 \leq x \leq \theta \\
&= 0 \text{ if, } x = \theta+n-N \\
&= \infty \text{ if, } x = \theta+1
\end{aligned}$$

which is non-decreasing in x. We can use the above to provide the following general version. Take $\theta_1 < \theta_2$

$$\frac{p(x,\theta_2)}{p(x,\theta_1)} = \frac{p(x,\theta_1+1)}{p(x,\theta_1)} \times \frac{p(x,\theta_1+2)}{p(x,\theta_1+1)} \times \cdots \times \frac{p(x,\theta_2)}{p(x,\theta_2-1)} \uparrow x.$$

Hence the family of Hypergeometric distribution has MLR in $T(x) = x$.

The following useful result is due to Karlin and Rubin (1956). We simply state the result without giving its proof.

Karlin-Rubin Theorem

For testing,

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta > \theta_0.$$

Corresponding to MLR family $\{p(x,\theta), \theta \in \Theta\}$ in $T(x)$, the test given by,

$$\varphi(x) = 1 \text{ if } T(x) > c$$

$$
\begin{aligned}
&= a \text{ if } T(x) = c \\
&= 0 \text{ if } T(x) < c
\end{aligned}
$$

with c and $a \in [0,1]$ are such that,

$$E_{\theta_0}\varphi(X) = P_{\theta_0}[T(X) > c] + aP_{\theta_0}[T(X) = c] = \alpha \in [0,1]$$

is UMP size α for (H, H_A).

Proof. Take any $\theta_1 > \theta_0$. Then by NP lemma, MP size α test for testing $H : \theta = \theta_0$ against $H'_A : \theta = \theta_1$is given by,

$$
\begin{aligned}
\varphi(x) &= 1 \text{ if } p(x,\theta_1) > kp(x,\theta_0) \\
&= a \text{ if } p(x,\theta_1) = kp(x,\theta_0) \\
&= 0 \text{ if } p(x,\theta_1) < kp(x,\theta_0)
\end{aligned}
$$

where $k \geq 0$ and $a \in [0,1]$ are such that,

$$E_{\theta_0}\varphi(X) = \alpha. \tag{1}$$

Let us write,

$$R(x) = \frac{p(x,\theta_1)}{p(x,\theta_0)} = g(T(x)),$$

where, $g(T(x))$ is a non-decreasing function of $T(x)$. Let us set,

$$c = inf\{T(x) : g(T(x)) \geq k\}$$

then the above test function can be written as,

$$
\begin{aligned}
\varphi(x) &= 1 \text{ if } T(x) > c \\
&= a \text{ if } T(x) = c \\
&= 0 \text{ if } T(x) < c
\end{aligned}
$$

and it satisfies (1), and equivalent to the desired test. Such a test being independent of $\theta_1 (> \theta_0)$, is UMP size α for

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta => \theta_0.$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Monotone Likelihood Ratio

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Result 1 The power function $\beta(\theta)$ of the test

$$\begin{aligned}\varphi(x) &= 1 \text{ if } T(x) > c \\ &= a \text{ if } T(x) = c \\ &= 0 \text{ if } T(x) < c\end{aligned}$$

is strictly increasing in θ so long as $0 < \beta(\theta) < 1$.

Proof We have seen that the test $\varphi(x)$ is UMP size α for testing

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta => \theta_0.$$

We have to show that $\beta(\theta_1) < \beta(\theta_2)$ provided $0 < \beta(\theta_1) < 1$ Note that $\varphi(x)$ is MP of size $\beta(\theta_1)$ for testing for testing $H^\star : \theta = \theta_1$ against $H_1^\star : \theta = \theta_2$. Since $0 < \beta(\theta_1) < 1$ the test $\varphi(x)$ is strictly unbiased i.e. $\beta(\theta_2) > \beta(\theta_1)$.

Corollary 1. Test given by,

$$\begin{aligned}\varphi(x) &= 1 \text{ if } T(x) > c \\ &= a \text{ if } T(x) = c \\ &= 0 \text{ if } T(x) < c\end{aligned}$$

is also UMP for,

$$H : \theta \leq \theta_0 \text{ vs. } H_A : \theta = \theta_1 (> \theta_0).$$

Proof. Since $\beta(\theta)$ is strictly increasing

$$\beta(\theta) \leq \beta(\theta_0) = \alpha \text{ for every } \theta : \theta \leq \theta_0.$$

Hence the test $\varphi(x)$ is also UMP for

$$H : \theta \leq \theta_0 \text{ vs. } H_A : \theta => \theta_0.$$

Result 2. A test $\varphi^\star(x)$ given by,

$$\begin{aligned}
\varphi^\star(x) &= 1 \text{ if } T(x) < c \\
&= a \text{ if } T(x) = c \\
&= 0 \text{ if } T(x) > c
\end{aligned}$$

with c and $a \in [0, 1]$ are such that,

$$E_{\theta_0}\varphi(X) = P_{\theta_0}[T(X) < c] + aP_{\theta_0}[T(X) = c] = \alpha \in [0, 1]$$

is UMP size α for

$$H : \theta \geq \theta_0 \text{ vs. } H_A : \theta < \theta_0$$

Proof For any $\theta_1 > \theta_0$ the ratio

$$R(x) = \frac{p(x, \theta_1)}{p(x, \theta_0)}$$

is a non-decreasing function of $T(x)$ i.e. a non-increasing function of $-T(x)$. If we replace θ_1 by $-\theta_1$ and θ_0 by $-\theta_0$ then $R(x)$ is a non-decreasing function of $-T(x)$. Hence the test $\varphi^\star(x)$ given by,

$$\begin{aligned}
\varphi^\star(x) &= 1 \text{ if } -T(x) > -c \\
&= a \text{ if } -T(x) = -c \\
&= 0 \text{ if } -T(x) < -c
\end{aligned}$$

is UMP size α for

$$H : -\theta \leq -\theta_0 \text{ vs. } H_A : -\theta > -\theta_0$$

Hence the result follows.

Result 3 The test

$$\begin{aligned}
\varphi(x) &= 1 \text{ if } T(x) > c \\
&= a \text{ if } T(x) = c \\
&= 0 \text{ if } T(x) < c
\end{aligned}$$

is Uniformly least powerful size α for testing $H : \theta = \theta_0$ vs. $H_A : \theta < \theta_0$

Proof. Consider the test $\varphi^c(x) = 1 - \varphi(x)$,

$$\begin{aligned}\varphi^c(x) &= 0 \text{ if } T(x) > c\\ &= a \text{ if } T(x) = c\\ &= 1 \text{ if } T(x) < c\end{aligned}$$

with $E_{\theta_0}(\varphi^c(X)) = 1 - \alpha$. Here $\varphi^c(x)$ is a left tailed test of size $1 - \alpha$. Such a test is UMP of size $1 - \alpha$ for testing $H : \theta = \theta_0$ vs. $H_A : \theta < \theta_0$ among all tests $\tilde{\varphi}(x) : E(\tilde{\varphi}(x)) \leq 1 - \alpha$. That is, $\varphi^c(x)$ satisfies,

$$E_\theta \varphi^c(X) \geq E_\theta \tilde{\varphi}(X)$$

for all $\theta < \theta_0$ along with $E_{\theta_0}\varphi^c(X) = 1 - \alpha$ and for all $\tilde{\varphi} : E_{\theta_0}\tilde{\varphi}(X) \leq 1 - \alpha$. That is,

$$E_\theta[1 - \varphi(X)] \geq E_\theta \tilde{\varphi}(X)$$

for all $\theta < \theta_0$ and for all $\tilde{\varphi} : E_{\theta_0}\tilde{\varphi}(X) = 1 - \alpha$, *i.e.* for all $\tilde{\varphi} : E_{\theta_0}(1 - \tilde{\varphi}(X)) = 1 - \alpha$. That is,

$$E_\theta[1 - \varphi(X)] \geq E_\theta[1 - \varphi^\star(X)]$$

for all $\theta < \theta_0$ and for all $\tilde{\varphi} : E_{\theta_0}\varphi^\star(X) = \alpha$. Hence we get $E_\theta\varphi(X) \leq E_\theta\varphi^\star(X)$ for all $\theta < \theta_0$ and $\varphi^\star : E_{\theta_0}\varphi^\star(X) = \alpha$. That means $\varphi(x)$ given by

$$\begin{aligned}\varphi(x) &= 1 \text{ if } T(x) > c\\ &= a \text{ if } T(x) = c\\ &= 0 \text{ if } T(x) < c\end{aligned}$$

is ULP size α within $\{\varphi^\star(x) : E_{\theta_0}\varphi^\star(X) = \alpha\}$.

Now we consider some applications of Karlin-Rubin Theorem. First we consider the some examples where the distribution belongs to the one-parameter exponential family and later we consider the distributions which do not belong to the exponential

family.

1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $N(\theta, 1)$, population. Since, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = \sum_{i=1}^{n} X_i$, a right tailed test based on $\sum_{i=1}^{n} X_i$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

2. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $N(0, \theta^2)$, population. Since, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = \sum_{i=1}^{n} X_i^2$, a right tailed test based on $\sum_{i=1}^{n} X_i^2$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

3. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $Bernoulli(\theta)$, population. Since, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = \sum_{i=1}^{n} X_i$, a right tailed test based on $\sum_{i=1}^{n} X_i$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

4. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $Geometric(\theta)$, population. Since, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = \sum_{i=1}^{n} X_i$, a right tailed test based on $\sum_{i=1}^{n} X_i$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

5. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from exponential population with mean $\frac{1}{\theta}$. Since, $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = -\sum_{i=1}^{n} X_i$, a left tailed test based on $\sum_{i=1}^{n} X_i$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

6. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from exponential population with mean θ. As we know that $\{p(x,\theta), \theta \in \Theta\}$ has MLR in $T(x) = \sum_{i=1}^{n} X_i$, a right tailed test based on $\sum_{i=1}^{n} X_i$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

7. Let X be an observation from $Cauchy(\theta, 1)$ population. We have seen that $Cauchy(\theta, 1)$ is not a member of MLR family so there does not exist any UMP test for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

8. Let X be an observation from $Cauchy(0, \theta)$ population. Since, $\{p(x, \theta), \theta \in \Theta\}$ has MLR in $T(x) = |X|$, a right tailed test based on $|X|$ is a UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

9. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $R(0, \theta)$ population. As we know that $\{p(x, \theta), \theta \in \Theta\}$ has MLR in $T(x) = X_{(n)}$, a right tailed test based on $X_{(n)}$ is UMP for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$.

At the end of our discussion we discuss the following interesting problem.

Problem : For $N(\theta, 1)$ population there does not exist UMP size α test for testing $H : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$.

Solution : Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a random sample from $N(\theta, 1)$, population. If possible suppose there exist UMP size α test for $H : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$ and let it be $\phi(\mathbf{X})$. Then the test $\phi(\mathbf{X})$ is UMP size alpha for testing $H : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$. Again, the test $\phi(\mathbf{X})$ is UMP size alpha for testing $H : \theta = \theta_0$ vs. $H_A : \theta < \theta_0$. This is contradiction to the fact that a test which is UMP for the right sided alternative becomes uniformly least powerful for left sided alternatives. Hence there does not exist UMP size α test for testing $H : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$.

Statistical Inference I

Generalized Neyman-Pearson lemma-Theory of UMPU tests

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In Modules 23 and 24 we have discussed the Neyman-Pearson (NP) lemma for obtaining most powerful tests. In the present module we shall consider a generalization of NP lemma. In this generalization we consider the same optimization problem considered in the NP lemma, but increase the number of side conditions from one to many. As in the case of NP lemma, we first state and prove the generalized NP (GNP) lemma for the non-randomized tests and then consider the case of randomized tests. In the npn-randomized case, the proof of the GNP lemma is very similar to the proof of the NP lemma.

epathshala पाठशाला A Gateway to All Post Graduate Courses

GNP Lemma (Non-Randomized case)

Let $g_0, g_1, \ldots, g_m$ be $(m+1)$ integrable real valued functions defined on $\mathcal{X}$. Suppose there exists at least one test function $\varphi(x)$ such that,

$$\int_{\mathcal{X}} \varphi(x) g_j(x) dx = c_j, j = 1, 2, \ldots, m, \tag{1}$$

where $c_1, \ldots, c_m$ are some known numbers. Let $\varphi_0(x)$ be another test function satisfying (1), such that,

$$\begin{aligned} \varphi_0(x) &= 1 \text{ if } g_0(x) > \sum_{j=1}^{m} k_j g_j(x) \\ &= 0 \text{ if } g_0(x) < \sum_{i=1}^{m} k_j g_j(x) \end{aligned} \tag{2}$$

where $k_1, \ldots, k_m$ are some constants determined appropriately. Then we have,

$$\int_{\mathcal{X}} \varphi_0(x) g_0(x) dx \geq \int_{\mathcal{X}} \varphi(x) g_0(x) dx \tag{3}$$

for all test $\varphi(x)$ satisfying (1).

Particular case

Suppose $m = 1, g_0(x) = p(x, \theta_1), g_1(x) = p(x, \theta_0), k_1 = k, c_1 = \alpha$. Then (1) is equivalent to the size condition i.e.

$$\int_{\mathcal{X}} \varphi(x) p(x, \theta_0) dx = \alpha$$

Then the test function $\varphi_0(x)$ the above equation and from (2) we see that

$$\begin{aligned} \varphi_0(x) &= 1 \text{ if } p(x, \theta_1) > kp(x, \theta_0) \\ &= 0 \text{ if } p(x, \theta_1) < kp(x, \theta_0) \end{aligned}$$

From equation (3) we get

$$\int_{\chi} \varphi_0(x) p(x, \theta_1) dx \geq \int_{\chi} \varphi(x) p(x, \theta_1) dx$$

Hence $\varphi_0(x)$ is an MP test of size α.

Proof

Let us define a function

$$Q(x) = (\varphi_0(x) - \varphi(x)) \left(g_0(x) - \sum_{i=1}^{m} k_j g_j(x) \right).$$

Then as in the proof of Neyman-Pearson lemma it can be shown that

$$Q(x) \geq 0 \text{ for all} x \in \mathcal{X}.$$

It follows that

$$\begin{aligned} \int Q(x) dx &\geq 0 \\ \Rightarrow \int_{\mathcal{X}} \varphi_0(x) g_0(x) dx - \int_{\mathcal{X}} \varphi(x) g_0(x) dx &\geq \int_{\mathcal{X}} \varphi_0(x) \sum_{j=1}^{m} k_j g_j(x) dx - \int_{\mathcal{X}} \varphi(x) \sum_{j=1}^{m} k_j g_j(x) dx \end{aligned}$$

epathshala पाठशाला
A Gateway to All Post Graduate Courses

$$
\begin{aligned}
&= \sum_{j=1}^{m} k_j \int_{\mathcal{X}} \varphi_0(x) g_j(x) g_j(x) dx - \sum_{j=1}^{m} k_j \int_{\mathcal{X}} \varphi(x) g_j(x) dx \\
&= \sum_{j=1}^{m} k_j c_j - \sum_{j=1}^{m} k_j c_j \\
&= 0
\end{aligned}
$$

Hence the proof.

GNP Lemma (Randomized case)

Let $g_0, g_1, \ldots, g_m$ be $(m+1)$ integrable real valued functions defined on $\mathcal{X}$. Suppose there exists at least one test function $\varphi(x)$ such that,

$$\int_{\mathcal{X}} \varphi(x) g_j(x) dx = c_j, j = 1, 2, \ldots, m, \tag{4}$$

where $c_1, \ldots, c_m$ are some known numbers. Let $\varphi_0(x)$ be another test function satisfying (4), such that,

$$
\begin{aligned}
\varphi_0(x) &= 1 \text{ if } g_0(x) > \sum_{j=1}^{m} k_j g_j(x) && (5) \\
&= \gamma(x) \text{ if } g_0(x) = \sum_{j=1}^{m} k_j g_j(x) && (6) \\
&= 0 \text{ if } g_0(x) < \sum_{i=1}^{m} k_j g_j(x) && (7)
\end{aligned}
$$

where $0 \le \gamma(x) \le 1$ and $k_1, \ldots, k_m$ are some constants determined appropriately. Then we have,

$$\int_{\mathcal{X}} \varphi_0(x) g_0(x) dx \ge \int_{\chi} \varphi(x) g_0(x) dx$$

for all test $\varphi(x)$ satisfying (4).

Unbiased test: A test φ is called unbiased test at level α for testing,

$$H : \theta \in \Theta_H \text{ vs. } H_A : \theta \in \Theta_{H_A} \subset (\Theta \cap \Theta_H^c),$$

if,

$$E_\theta \varphi \le \alpha \ \ \forall \theta \in \Theta_H \ \text{ and } \ E_\theta \varphi \ge \alpha \ \ \forall \theta \in \Theta_{H_A} \tag{8}$$

In Module 32 we have already seen that UMP does not exist for testing $H : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$ where the sample is taken from a $N(\theta, 1)$ population. In such a situation, we may impose the restriction of unbiasedness and look for an U.M.P. test in the class of all unbiased tests.

Uniformly most powerful unbiased test (UMPU): A test φ^0 is called UMPU test at level α for testing,

$$H : \theta \in \Theta_H \text{ vs. } H_A : \theta \in \Theta_{H_A} \subset (\Theta \cap \Theta_H^c),$$

if,

$$(i) E_\theta \varphi^0 \leq \alpha \;\; \forall \theta \in \Theta_H (\text{Size condition})$$
$$(ii) E_\theta \varphi^0 \geq \alpha \;\; \forall \theta \in \Theta_{H_A} (\text{Unbiased condition})$$
$$(iii) E_\theta \varphi^0 \geq E_\theta \varphi \;\; \forall \theta \in \Theta_{H_A} (\text{Power condition}),$$

where φ is any test function satisfying (i) and (ii).

Note If $\beta_\varphi(\theta) = E_\theta \varphi$ is a continuous function of θ, then (8) implies

$$\beta_\varphi(\theta) = \alpha \text{ for all } \theta \in \Theta_B \qquad (9),$$

where Θ_B is the common boundary of Θ_H and Θ_{H_A}, that is, the set of points θ that are limit points of both Θ_H and Θ_{H_A}. Tests satisfying (9) are said to be α similar on the boundary.

Example . Let $X_1, X_2, \ldots, X_n \sim N(\theta, 1)$, θ is unknown. To test,

$$H : \theta \leq 0 \text{ vs. } H_A : \theta > 0.$$

Consider the test

$$\varphi(\mathbf{x}) = 1 \text{ if, } \sqrt{n}\bar{x} > \tau_\alpha$$
$$0 \text{ if, } \sqrt{n}\bar{x} \leq \tau_\alpha$$

where τ_α is the upper $100\alpha\%$ point of $N(0,1)$ distribution.

Here,

$$\Theta_H = \{\theta : \theta \leq 0\},$$

$$\Theta_{H_A} = \{(\theta : \theta > 0\}$$

and,

$$\Theta_B = \{\theta : \theta = 0\}.$$

Note that on Θ_B,

$$\sqrt{n}\bar{x} \sim N(0,1)$$

Therefore,

$$\begin{aligned} E_\theta \varphi &= P\left[\sqrt{n}\bar{x} > \tau_\alpha\right] \\ &= \alpha, \forall \theta \in \Theta_B. \end{aligned}$$

Hence, φ is $\alpha-$similar on Θ_B.

epathshala पाठशाला A Gateway to All Post Graduate Courses

Since it is more convenient to work with (9) than with (8), the following result plays an important role in the determination of UMPU tests.The proof of the result will be discussed later.

Result 1. If power function of a test φ is continuous, then unbiasedness of a level α test φ implies, its $\alpha-$similarity.

Tests for One Parameter Exponential family of distributions We have already seen that in case of one parameter exponential family of distributions one can get UMP test for (i) $H : \theta \leq \theta_0$ against $H_A : \theta > \theta_0$ by applying Karlin-Rubin theorem. In can be shown that (see Lehmann) there exists UMP test for (ii) $H : \theta \leq \theta_1$ or $\theta \geq \theta_2 (\theta_1 < \theta_2)$ against $H_A : \theta_1 < \theta < \theta_2$ by using GNP lemma. If $T(x)$ is sufficient for θ

then the UMP test of size α for testing problem (ii) is given by

$$\begin{aligned}\varphi_0(x) &= 1 \text{ if } k_1 < T(x) < k_2 \\ &= \gamma_i \text{ if } T(x) = k_i, i = 1, 2, \\ &= 0 \text{ if } T(x) < k_1 \text{ or } T(x) > k_2,\end{aligned}$$

where $k's$ and $\gamma's$ are determined by

$$E_{\theta_1}\varphi(X) = E_{\theta_2}\varphi(X) = \alpha$$

.

But there does not exist UMP test for (iii) $H : \theta_1 \leq \theta \leq \theta_2$ against $H_A : \theta < \theta_1$ or $\theta > \theta_2$ and for (iv) $H : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$. For the testing problems (iii) and (iv) we can find UMPU tests by using the GNP lemma. We shall now consider the testing problem (iv). To find a UMPU test for (iv) let us define

$$U(\alpha, \theta_0) = \{\varphi; E_{\theta_0}\varphi \leq \alpha \text{ and, } E_\theta\varphi \geq \alpha, \forall\theta \neq \theta_0\},$$

class of unbiased test at level α for testing the hypothesis under study. It can be shown that (see Lehamnn) the power function $E_\theta\varphi$ of any test $\varphi(X)$ for the one parameter exponential family is continuous and differentiable. So the unbiasedness condition (8) implies $E_\theta\varphi$ has minimum value at $\theta = \theta_0$ and hence $E_{\theta_0}\varphi = \alpha$ and $E'_{\theta_0}\varphi = 0$, where,

$$E'_{\theta_0}\varphi = \frac{d}{d\theta}E_\theta\varphi|_{\theta=\theta_0}.$$

Now we define,

$$D(\alpha, \theta_0) = \left\{\varphi; E_{\theta_0}\varphi = \alpha, \text{ and, } E'_{\theta_0}\varphi = 0\right\}.$$

Result 2. Under the assumption $E_\theta\varphi$ is continuous, $U(\alpha, \theta_0) \subset D(\alpha, \theta_0)$.

Proof:

$$\varphi \in U(\alpha, \theta_0) \Leftrightarrow E_{\theta_0}\varphi \leq \alpha \text{ and, } E_\theta\varphi \geq \alpha, \forall\theta \neq \theta_0\}.$$

Therefore,

$$E_{\theta_0}\varphi = \alpha, \text{ and, } E'_{\theta_0}\varphi = 0.$$

Hence, $\varphi \in D(\alpha, \theta_0)$.

Result 3. Suppose φ^0 is the best test within $D(\alpha, \theta_0)$ that is φ^0 maximizes $E_{\theta_1}\varphi^0 \ \forall\theta_1 \neq \theta_0$. s.t.,

$$E_{\theta_0}\varphi = \alpha \text{ and } E'_{\theta_0}\varphi = 0.$$

Then, $\varphi^0 \in U(\alpha, \theta_0)$.

Proof: Let,

$$\begin{aligned} \varphi^\star(\mathbf{x}) &= \alpha, \quad \forall\mathbf{x} \\ E_{\theta_0}\varphi^\star &= \alpha \\ E_{\theta_1}\varphi^\star &= \alpha \\ E'_{\theta_0}\varphi^\star &= 0. \end{aligned}$$

Therefore, $\varphi^\star \in D(\alpha, \theta_0)$. Hence, $E_{\theta_1}\varphi^0 \geq E_{\theta_1}\varphi^\star = \alpha, \forall\theta_1 \neq \theta_0$. Also, $E_{\theta_0}\varphi^0 = \alpha \Rightarrow E_{\theta_1}\varphi \geq E_{\theta_0}\varphi^0, \forall\theta_1 \neq \theta_0$. Hence $\varphi^0 \in U(\alpha, \theta_0)$ and it is UMPU at level α for testing,

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta \neq \theta_0.$$

Remark From Result (3) we see that in case of a OPEF a UMPU test for $H : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$ can be obtained by determining a UMP test within the class $D(\alpha, \theta_0)$.

Statistical Inference I
Locally most powerful tests

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

So far we have treated the testing of one-sided and two-sided problems for a real parameter when the distribution of the random observable X is sufficiently well behaved, that is, a one-parameter exponential family or a family with monotone likelihood ratio. In this module we consider the problem of finding optimal tests for distributions without a monotone likelihood ratio and which does not belong to one-parameter exponential family. Among the various other methods of available in the literature we are going to discuss the locally best tests of Neyman and Pearson.

One sided tests

Let X be a random variable with p.m.f/p.d.f. $p_\theta(x), \theta \in \Theta \subset \mathcal{R}$.

Consider the problem of testing,

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta > \theta_0$$

Definition. A test φ^0 is called a Locally most powerful (LMP) test at size α for testing,

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta > \theta_0$$

if for some $\epsilon > 0$,

$$(i) \quad \beta_{\varphi_0}(\theta_0) = \alpha$$

$$(ii) \quad \beta_{\varphi_0}(\theta) \geq \beta_\varphi(\theta) \text{ for all } \theta \in (\theta_0, \theta_0 + \epsilon) \text{ and whatever } \varphi \text{ satisfying } (i)$$

where $\beta_\varphi(\theta) = E_\theta \varphi$ denotes the power function of the testφ. By Mean value theorem

$$\beta_\varphi(\theta) = \beta_\varphi(\theta_0) + (\theta - \theta_0)\beta'_\varphi(\theta^\star), \;\; \theta_0 < \theta^\star < \theta.$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Assumptions

(i) $\beta_\varphi(\theta)$ is continuously differentiable in the neighborhood of θ_0 for every φ.

(ii) $\beta'_\varphi(\theta) = \int \varphi \frac{\partial p_\theta(x)}{\partial \theta} dx.$

Our problem is to minimize $\beta'_\varphi(\theta_0)$ subject to, $\beta_\varphi(\theta_0) = \int \varphi p_{\theta_0}(x)dx = \alpha$. Using generalized NP Lemma we get the optimum choice of φ as,

$$\begin{aligned}
\varphi^0 &= 1 \text{ if } \frac{\partial p_\theta(x)}{\partial \theta_0} > k p_{\theta_0}(x) \\
&= \gamma(x) \text{ if } \frac{\partial p_\theta(x)}{\partial \theta_0} = k p_{\theta_0}(x) \\
&= 0 \text{ if } \frac{\partial p_\theta(x)}{\partial \theta_0} < k p_{\theta_0}(x)
\end{aligned}$$

$\Leftrightarrow$

$$\begin{aligned}
\varphi^0 &= 1 \text{ if } \frac{\partial \log p_\theta(x)}{\partial \theta_0} > k \\
&= \gamma(x) \text{ if } \frac{\partial \log p_\theta(x)}{\partial \theta_0} = k \\
&= 0 \text{ if } \frac{\partial \log p_\theta(x)}{\partial \theta_0} < k
\end{aligned}$$

where, k and γ are such that, $E_{\theta_0}\varphi = \alpha$.

Example. Let $X_1, X_2 \ldots, X_n \sim$ Cauchy(θ,1). The joint density is given by,

$$p_\theta(\mathbf{x}) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{(1 + (x_i - \theta)^2)}.$$

Consider the problem of testing,

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta > \theta_0$$

LMP test given by,

$$\begin{aligned}
\varphi^0(\mathbf{x}) &= 1 \text{ if } \frac{\partial \log p_\theta(x)}{\partial \theta_0} > k \\
&= 0 \text{ if } \frac{\partial \log p_\theta(x)}{\partial \theta_0} < k.
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial \log p_\theta(x)}{\partial \theta_0}\\
&= \frac{\partial}{\partial \theta_0}\left(\text{const.} \sum_{i=1}^{n} \log\left[\frac{1}{1+(x_i-\theta)^2}\right]\right)\\
&= \sum_{i=1}^{n}\left[\frac{2(x_i-\theta)}{1+(x_i-\theta)^2}\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\varphi^0(\mathbf{x}) = 1 &\text{ if } 2\sum_{i=1}^{n}\left[\frac{x_i}{1+x_i^2}\right] > c\\
0 &\text{ if } 2\sum_{i=1}^{n}\left[\frac{x_i}{1+x_i^2}\right] < c.
\end{aligned}
$$

under $H: \theta = 0$ *i.e.*, $H: X \sim C(0,1)$. So,

$$E\left(\frac{2X}{1+X^2}\right) = 0$$

and,

$$V\left(\frac{2X}{1+X^2}\right) = \frac{1}{2},$$

since, X has a symmetric distribution about 0. So, by CLT we have,

$$\frac{2\sum_{i=1}^{n}\frac{X_i}{1+X_i^2}}{\sqrt{n/2}} \Rightarrow N(0,1).$$

Hence for large enough n, we can choose, $c = \tau_\alpha\sqrt{n/2}$.

<u>Drawback.</u> If $\alpha < 1/2$, for $c > 0$ $\left\{\mathbf{x}; 2\sum_{i=1}^{n}\frac{x_i}{1+x_i^2} > c\right\}$ is a bounded set *i.e.*,

$$\left\{\mathbf{x}; 2\sum_{i=1}^{n}\frac{x_i}{1+x_i^2} > c\right\} \subseteq \left\{\mathbf{x}; \sum_{i=1}^{n} x_i^2 < R^2\right\}, 0 < R^2 < \infty.$$

Therefore, the power function,

$$
\begin{aligned}
P_\theta\left[2\sum_{i=1}^{n}\frac{X_i}{1+X_i^2} > c\right] &\leq \int_{\sum_{i=1}^{n} x_i^2 < R^2} \frac{1}{\pi^n}\prod_{i=1}^{n}\frac{1}{(1+(x_i-\theta)^2)}d\mathbf{x}\\
&\leq \int_{|x_1|<R}\frac{1}{\pi}\frac{1}{(1+(x_1-\theta)^2)}dx_1 \text{ since } \sum_{i=1}^{n} x_i^2 < R^2 \Rightarrow |x_1| < R\\
&= \int_{-R-\theta}^{R-\theta}\frac{1}{\pi}\frac{1}{(1+z^2)}dz \longrightarrow 0 \text{ as } \theta \longrightarrow \infty.
\end{aligned}
$$

Thus the power goes to 0 as $\theta \longrightarrow \infty$. Hence $\varphi^0(\mathbf{x})$ is good at detecting small departures from the null hypothesis, it is unsuccessful in detecting sufficiently large ones.

Two sided tests

The notion of a locally best test may be extended to testing the hypothesis

$$H : \theta = \theta_0 \text{ vs. } H_A : \theta \neq \theta_0$$

by specifying the slope of the power function at θ_0 and requiring the second derivative of the power function at θ_0 to be a maximum. Hence we consider only unbiased tests with slope zero at θ_0 .

Assumptions

(i) $\beta_\varphi(\theta)$ is twice continuously differentiable in the neighbourhood of θ_0 for every φ.

(ii) $\beta'_\varphi(\theta) = \int \varphi \frac{\partial p_\theta(x)}{\partial \theta} dx.$

(iii) $\beta''_\varphi(\theta) = \int \varphi \frac{\partial^2 p_\theta(x)}{\partial \theta^2} dx.$

By Mean value theorem

$$\beta_\varphi(\theta) = \beta_\varphi(\theta_0) + (\theta - \theta_0)\beta'_\varphi(\theta_0) + \frac{(\theta - \theta_0)^2}{2!}\beta'_\varphi(\theta^\star), \;\; min(\theta_0, \theta) < \theta^\star < max(\theta_0, \theta).$$

Locally best unbiased tests

A test φ^0 is called locally best unbiased test at size α for testing,

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

if out of all φ satisfying $\beta_\varphi(\theta_0) = \alpha$ and $\beta'_\varphi(\theta_0) = 0$ the test φ^0 maximizes the value of the second derivative at θ_0 , that is,

$$\beta''_{\varphi^0}(\theta_0) \geq \beta'_\varphi(\theta_0)$$

. 'When such a test. is unique, the optimum property of a locally best un- biased test may be stated as follows:

A test φ^0 is called LMP test at size α for testing,

$$H_{:}\theta = \theta_0 \text{ vs. } H_A : \theta \neq \theta_0$$

if for some $\epsilon > 0$,

$$\begin{aligned} &(i) \quad \beta_{\varphi_0}(\theta_0) = \alpha \\ &(ii) \quad \beta_{\varphi_0}(\theta) \geq \beta_{\varphi}(\theta), \quad \forall |\theta - \theta_0| < \epsilon, \theta \neq \theta_0 \end{aligned}$$

where, $\beta_{\varphi}(\theta)$ is the power function of the test φ and φ satisfies (i). Our aim is to minimize

$$\int \varphi \frac{\partial^2 p_\theta(x)}{\partial \theta^2}|_{\theta=\theta_0} dx$$

subject to,

$$\int \varphi p_\theta(x) dx = \alpha,$$

and

$$\int \varphi \frac{\partial p_\theta(x)}{\partial \theta}|_{\theta=\theta_0} dx = 0,$$

$$\begin{aligned} \varphi_0(x) = 1 \text{ if } & \frac{\partial^2 p_\theta(x)}{\partial \theta^2}|_{\theta=\theta_0} > k_1 p_{\theta_0}(x) + k_2 \frac{\partial p_\theta(x)}{\partial \theta}|_{\theta=\theta_0} \\ 0 \text{ if } & \frac{\partial^2 p_\theta(x)}{\partial \theta^2}|_{\theta=\theta_0} < k_1 p_{\theta_0}(x) + k_2 \frac{\partial p_\theta(x)}{\partial \theta}|_{\theta=\theta_0}. \end{aligned}$$

and, $E_{\theta_0}\varphi^0 = \alpha$ and $E'_{\theta_0}\varphi^0 = 0$. Note that,

$$\frac{\partial \log p_\theta(x)}{\partial \theta} = \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta}$$

and,

$$\frac{\partial^2 \log p_\theta(x)}{\partial \theta^2} = -\frac{1}{(p_\theta(x))^2} \left[p_\theta(x) \frac{\partial^2 p_\theta(x)}{\partial \theta^2} - \left(\frac{\partial p_\theta(x)}{\partial \theta} \right)^2 \right]$$

Therefore, $\varphi^0(x)$ can be rewritten as,

$$\varphi_0(x) = 1 \text{ if } \frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}|_{\theta=\theta_0} + \left(\frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=\theta_0}\right)^2 > k_1 + k_2 \frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=\theta_0}$$
$$0 \text{ if } \frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}|_{\theta=\theta_0} + \left(\frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=\theta_0}\right)^2 < k_1 + k_2 \frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=\theta_0}.$$

Example. Let $X_1, \ldots, X_n \sim C(\theta, 1)$ To test,

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta \neq 0$$

Since, $X_i \sim C(\theta, 1)$, the joint density is given by,

$$p_\theta(x) = \frac{1}{\pi^n \prod_{i=1}^n (1 + (x_i - \theta)^2)}$$

and also,

$$\frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=0} = \sum_{i=1}^n \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}|_{\theta=0} = \sum_{i=1}^n \frac{2x_i}{1 + x_i^2}$$
$$\frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}|_{\theta=0} = 2\sum_{i=1}^n \frac{-[1 + (x_i - \theta)^2] + 2(x_i - \theta)}{[1 + (x_i - \theta)^2]^2}|_{\theta=0} = 2\sum_{i=1}^n \frac{2x_i^2 - 1}{(1 + x_i^2)^2}.$$

Therefore, LMP test is given by,

$$\varphi_0(x) = 1 \text{ if } v > k_1 + k_2 u$$
$$0 \text{ if } v < k_1 + k_2 u.$$

where, $u = \sum_{i=1}^n \frac{2x_i}{1+x_i^2}$, and $v = 2\sum_{i=1}^n \frac{2x_i^2-1}{(1+x_i^2)^2} + u^2$. Also, k_1, k_2 are such that, $E_{\theta=0}\varphi^0 = \alpha$, and $E'_{\theta=0}\varphi^0 = 0$. Observe that, $E_\theta\varphi = \int \varphi p_\theta(x)dx$. Hence,

$$\frac{\partial E_\theta \varphi}{\partial\theta}|_{\theta=0} = \int \varphi \frac{\partial \log p_\theta(x)}{\partial\theta}|_{\theta=0} p_\theta(x),$$

and

$$p'_\varphi(\theta_0) = 0 \Leftrightarrow E_{\theta_0}\varphi U = 0.$$

Now,

$$\begin{aligned}
E_{\theta_0}\varphi^0 U &= E_{\theta_0} U I\left(V > k_1 + k_2 U\right) \\
&= E_{\theta_0} U I\left(U > 0, V > k_1 + k_2 U\right) + E_{\theta_0} U I\left(U < 0, V > k_1 + k_2 U\right) \\
&= E_{\theta_0} U I\left(U > 0, V > k_1 + k_2 U\right) - E_{\theta_0} U I\left(U > 0, V > k_1 - k_2 U\right) \\
&= -E_{\theta_0} U I\left(U > 0, k_1 - k_2 U < V < k_1 + k_2 U\right).
\end{aligned}$$

Therefore, $E_{\theta_0}\varphi^0 U = 0 \Leftrightarrow k_2 = 0$. Hence the above test is given by,

$$\varphi_0(x) = 1 \text{ if } v > k_1$$
$$0 \text{ if } v < k_1.$$

where k_1 is such that, $E_{\theta_0}\varphi^0 = \alpha$.

ePathshala पाठशाला
A Gateway to All Post Graduate Courses

Statistical Inference I
UMPU tests for multi-parameter exponential family-I

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

We have already discussed the construction of UMPU test for the distributions belonging to one parameter exponential family. In case of multi-parameter exponential family the number of parameters involved is more than one, and we are concerned with testing a hypothesis relating to one but not all of them, the other parameters left unspecified by the hypothesis. Such parameters are called nuisance parameters. For example, in the testing for the mean of a normal population, the unknown variance is a nuisance a parameter. One method of overcoming the problem of nuisance parameter in hypothesis testing is to use similar tests which have been discussed in Module-33. In the present module, first we are going to re-introduce the concept of similar tests and then we shall introduce another important concept, viz., tests with Neyman structure.

Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p_\theta(\mathbf{x})$ with $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \mathcal{R}^k,\ k \geq 2$. Our problem is to test,

$$H : \theta \in \Theta_H \text{ vs. } H_A : \theta \in \Theta_{H_A} \left(\subseteq \Theta - \Theta_H\right).$$

where Θ_B is the set of all points which are limit points of both Θ_H and Θ_{H_A} i.e. $\Theta_B = \bar{\Theta}_H \cap \bar{\Theta}_{H_A}$, $\bar{A}$ is the derived set of A. It is the set of all points on the common boundary of Θ_H and Θ_{H_A}.

Similar test:

A test φ is called $\alpha-$similar on Θ_B if $E_\theta \varphi = \alpha,\ \forall \theta \in \Theta_B$.

Example 1. Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, $\mu.\sigma^2$ are unknown. Here $\theta = (\theta_1, \theta_2) = (\mu, \sigma), k = 2$. To test,

$$H : \mu \leq 0 \text{ vs. } H_A : \mu > 0.$$

$$\begin{aligned}\varphi(\mathbf{x}) &= 1 \text{ if, } \frac{\sqrt{n}\bar{x}}{s} > t_{\alpha;n-1} \\ &= 0 \text{ if, } \frac{\sqrt{n}\bar{x}}{s} \leq t_{\alpha;n-1}\end{aligned}$$

Here,

$$\Theta_H = \{(\mu, \sigma); \mu \leq 0\},$$

$$\Theta_{H_A} = \{(\mu, \sigma); \mu > 0\},$$

and,

$$\Theta_B = \{(\mu, \sigma); \mu = 0\}.$$

Note that on Θ_B,

$$\frac{\sqrt{n}\bar{x}}{\sigma} \sim N(0,1)\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

are independent. Therefore,

$$\frac{\sqrt{n}\bar{x}}{s} \sim t_{n-1},$$

and

$$\begin{aligned}E_\theta\varphi &= P\left[\frac{\sqrt{n}\bar{x}}{s} > t_{\alpha;n-1}\right] \\ &= \alpha, \forall\theta \in \Theta_B. \qquad (1)\end{aligned}$$

Hence, φ is $\alpha-$similar on Θ_B.

Result 1. If power function of a test φ is continuous, then unbiasedness of a level α test φ implies, its $\alpha-$similarity.

Proof. Suppose, $\theta^\star \in \Theta_B$,. Then there exists a sequence $\{\theta_n\}$, $\theta_n \in \Theta_H$, such that $\theta_n \longrightarrow \theta^\star$. Since $E_\theta(\varphi)$ is continuous, $E_{\theta_n}(\varphi) \longrightarrow E_{\theta^\star}(\varphi)$; and since $E_{\theta_n}(\varphi) \leq \alpha$ for $\theta_n \in \Theta_H$, $E_{\theta^\star}(\varphi) \leq \alpha$. Similarly, there exists a sequence $\tilde{\theta}_n$, $\theta_n \in \Theta_{H_A}$ such that $E_{\tilde{\theta}_n}(\varphi) \geq \alpha$ (since φ is unbiased) and $\tilde{\theta}_n \longrightarrow \theta^\star$. Thus $E_{\tilde{\theta}_n}(\varphi) \longrightarrow E_{\theta^\star}(\varphi))$, and it follows that $E_{\theta^\star}(\varphi) \geq \alpha$. Hence $E_{\theta^\star}\varphi = \alpha$, for all $\theta^\star \in \Theta_B$ and φ is α-similar on Θ_B.

Result 2. If power function of every test is continuous and level α test φ^0 is UMP within in the class of $\alpha-$similar test on Θ_B for testing $H : \theta \in \Theta_H$ against $H_A : \theta \in \Theta_{H_A}$, it is UMPU at level α for testing the same hypothesis.

Proof. Define,

$$\mathcal{C}_S = \{\varphi; E_\theta\varphi = \alpha, \forall\theta \in \Theta_B\} = \text{ class of all } \alpha \text{ similar tests on } \Theta_B,$$

and

$$\mathcal{C}_U = \{\varphi; E_\theta\varphi \leq \alpha, \forall\theta \in \Theta_H, E_\theta\varphi \geq \alpha, \forall\theta \in \Theta_{H_A}\} = \text{ class of all unbiased level } \alpha \text{ tests.}$$

From Result- 1, we get $\mathcal{C}_U \subseteq \mathcal{C}_S$, φ^0 be UMP with $\mathcal{C}_S$. Therefore for every $\theta \in \Theta_B$, there exists a θ_n such that, $\theta_n(\in \Theta_{H_A}) \longrightarrow \theta \in \Theta_B$, since θ is a cluster point of Θ_{H_A},

$$\Rightarrow E_{\theta_n}\varphi^0 \longrightarrow E_\theta\varphi^0 = \alpha$$
$$\Rightarrow E_{\theta_n}\varphi^0 \longrightarrow \alpha.$$

Consider, $\varphi^\star = \alpha, \forall\mathbf{x}$,

$$\begin{aligned} E_\theta\varphi^\star &= \alpha \;\forall\theta \in \Theta_B \Rightarrow \varphi^\star \in \mathcal{C}_S \\ E_\theta\varphi^0 &\geq E_\theta\varphi^\star = \alpha, \forall\theta \in \Theta_{H_A} \end{aligned}$$

Hence, φ^0 is a UMPU test.

Suppose T is sufficient for $\theta \in \Theta$, then for any test φ, we have

$$\psi(t) = E\left[\varphi(X)|t\right]$$

is independent of θ and it belongs to $[0, 1]$. Hence it would be natural to consider test based on $T^\star$ only instead of X. This fact motivates us to introduce a new concept, viz., tests with Neyman structure, which is defined as follows.

Tests with Neyman structure:

Let T be sufficient for Θ_B. A test φ is said to have Neyman structure at level α with respect T over Θ_B if,

$$E(\varphi|T) = \alpha \text{ a.e. } \mathcal{P}_B^T,$$

where,

$$\mathcal{P} = \{P_\theta; \theta \in \Theta\},$$

$$\mathcal{P}^T = \left\{P_\theta^T; \theta \in \Theta\right\},$$

and

$$\mathcal{P}_B^T = \left\{P_\theta^T; \theta \in \Theta_B\right\}.$$

Also, $E_\theta(\varphi|T) = \alpha$ on a set A, where $P_\theta^T(A) = 0$ and $P_\theta^T \in \mathcal{P}_B^T$.

Result 3. If φ has Neyman structure at level α with respect T, then it is $\alpha-$similar.

Proof. Note that,

$$\begin{aligned} E_\theta\varphi &= E_\theta E\left[\varphi|T\right] \\ &= \alpha, \ \forall\theta \in \Theta_B. \end{aligned}$$

The advantage of tests having Neyman structure is that on each of the surfaces $T = t$, the distribution of $\mathbf{X}$ is independent of the parameter θ. This provides a way of getting rid of the nuisance parameters. The following result provides an answer to the question 'When will similar tests have Neyman structure?'.

Result 4. Let T be a sufficient statistic for $\theta \in \Theta_B$. A necessary and sufficient condition for every $\alpha-$similar test has Neyman structure at level α with respect to T is the bounded completeness of $\mathcal{P}_B^T$.

Proof. (If part): φ is a $\alpha-$similar on Θ_B,

$$\Rightarrow E_\theta\varphi = \alpha \ \forall\theta \in \Theta_B$$

$$\Rightarrow E_\theta\left[E(\varphi|T) - \alpha\right] = 0 \ \forall\theta \in \Theta_B$$

$$\Rightarrow E(\varphi|T) = \alpha \text{ a.e. } \mathcal{P}_B^T,$$

since $\mathcal{P}_B^T$ is boundedly complete.

Proof. (Only if part): If possible, $\mathcal{P}_B^T$ is not boundedly complete. Therefore, there

exists a bounded function $\psi(t)$ and $P_B^\star \in \mathcal{P}_B^T$ such that, $E_\theta\psi(T) = 0$, $P_B^\star(\psi(T) \neq 0) > 0$,. Since $\psi(T)$ is a bounded function there exists a positive real number M such that $|\psi(T)| \leq M < \infty$. Define, $\varphi(t) = \alpha + c\psi(t), c = \frac{Min(\alpha, 1-\alpha)}{M}$.

Then $-cM + \alpha \leq \varphi(t) \leq cM + \alpha$.

Now

$$\begin{aligned} cM + \alpha &= 1 \text{ if } \alpha > \frac{1}{2} \\ &= 2\alpha \text{ if } \alpha < \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} -cM + \alpha &= 2\alpha - 1 \text{ if } \alpha > \frac{1}{2} \\ &= 0 \text{ if } \alpha < \frac{1}{2} \end{aligned}$$

Thus $\varphi(t)$ is a test function i.e. $0 \leq \varphi(t) \leq 1$.

Now

$$\begin{aligned} E_\theta\varphi(T) &= cE_\theta\psi(T) + \alpha \\ &= \alpha \text{ for all } \theta \in \Theta_B \end{aligned}$$

Hence $\varphi(T)$ is α-similar on Θ_B.

But

$$\begin{aligned} E_\theta\left(\varphi(T)|T = t\right) &= \varphi(t) \\ &= \alpha + c\psi(t) \\ &\neq \alpha \text{ with positive probability } P_B^\star \in \mathcal{P}_B^T \end{aligned}$$

So φ does not have a Neyman structure. Which is a contradiction. Hence $\mathcal{P}_B^T$ must be boundedly complete.

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
UMPU tests for multi-parameter exponential family-II

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In the last module we have shown that if power function of every test is continuous then UMP test φ^0 within in the class of $\alpha-$similar test is UMPU at level α for testing $H : \theta \in \Theta_H$ vs. $H_A : \theta \in \Theta_{H_A}$. In the present module we are going to explore this idea for constructing UMPU test when the distribution of $\mathbf{X}$ belonging to the k-parameter exponential family.

The p.m.f/p.d.f. of $\mathbf{X}$ is given by,

$$p_\theta(\mathbf{x}) = A(\theta) \exp^{\sum_{i=1}^{k} \theta_i T_i(\mathbf{x})} h(\mathbf{x}),$$

where $\theta = (\theta_1, \ldots, \theta_k) = (\theta_1, \theta^{(2)})$, $\mathbf{T} = (T_1, \ldots, T_k) = (T_1, T^{(2)}), k \geq 2$. Since we are dealing with MPEF we can consider the following set of assumptions,

(i) Power function of any test φ is continuous.

$(ii) \dfrac{\partial^m}{\partial\theta^m} \int \varphi p_\theta(x) d\mu = \int \varphi \dfrac{\partial^m}{\partial\theta^m} p_\theta(x) d\mu$

(iii) Conditional distribution of $T_1 | T^{(2)} = t^{(2)}$

$= \dfrac{A(\theta) \exp^{\theta_1 t_1 + \sum_{i=2}^{k} \theta_i t_i} h(\mathbf{t})}{A(\theta) \exp^{\sum_{i=2}^{k} \theta_i t_i} \int \exp^{\theta_1 t_1} h(\mathbf{t}) dt_1}$

$= A(\theta, t^{(2)}) \exp^{\theta_1 t_1} H^\star(t_1, t^{(2)})$

$=$ One Parameter Exponential Family (OPEF) .

$(iv) T^{(2)}$ is complete sufficient for $\mathcal{P}_B^T$ provided Θ_B contains a $(k-1)$ dimensional rectangle.

First we consider the problem of testing,

$$(I) H : \theta_1 \leq \theta_{10} \text{ vs. } H_A : \theta_1 > \theta_{10},$$

Since the distribution of $T_1|T^{(2)} = t^{(2)}$ belonging to OPEF, it possess the MLR property. By Karlin-Rubin theorem UMP test for testing $H : \theta_1 = \theta_{10}$ vs. $H_A : \theta_1 > \theta_{10}$, at level α based on the conditional distribution of $T_1|T^{(2)} = t^{(2)}$ is given by,

$$\begin{aligned}
\varphi^0(t_1, t^{(2)}) &= 1 \text{ if } t_1 > c(t^{(2)}) \\
&= \gamma(t^{(2)}) \text{ if } t_1 = c(t^{(2)}) \\
&= 0 \text{ if } t_1 < c(t^{(2)})
\end{aligned}$$

where $c(t^{(2)})$ and $\gamma(t^{(2)})$ are such that $E_{\theta_{10}}\left[\varphi^0(t_1, t^{(2)})|T^{(2)} = t^{(2)}\right] = \alpha$ a.e. $\mathcal{P}_B^{T^{(2)}}$.
Now we have to show that
(i) $\varphi^0(t_1, t^{(2)})$ is UMP within the class of all α-similar tests on $\Theta_B = \{\theta : \theta_1 = \theta_{10}\}$
(ii) $\varphi^0(t_1, t^{(2)})$ is a level α test for testing $H : \theta_1 \leq \theta_{10}$ vs. $H_A : \theta_1 > \theta_{10}$,.
Let us define

$$\mathcal{C}_{NS} = \left\{\varphi : E\varphi|T^{(2)} = \alpha \text{ a.e. } \mathcal{P}_B^{T^{(2)}}\right\},$$

$$\mathcal{C}_S = \{\varphi : E_\theta\varphi = \alpha\forall\theta \in \Theta_B\},$$

and,

$$\mathcal{C}_U = \{\varphi : E_\theta\varphi \leq \alpha\forall\theta \in \Theta_H, E_\theta\varphi \geq \alpha\forall\theta \in \Theta_{H_A}\}.$$

Now $\varphi^0 \in \mathcal{C}_{NS}$ and hence $\varphi^0 \in \mathcal{C}_S$. Let φ be any $\alpha-$similar test on Θ_B. Since $T^{(2)}$ is complete (hence boundedly complete) for Θ_B. Then any α similar has a Neyman structure by result 4, *i.e.*,

$$E\varphi|T^{(2)} = \alpha \text{ a.e. } \mathcal{P}_B^{T^{(2)}}.$$

Since φ^0 is UMP within $\mathcal{C}_{NS}$,

$$\begin{aligned}
E_\theta\varphi^0 &= E_{\theta^{(2)}}E_{\theta_1}\varphi^0|T^{(2)} \\
&\geq E_{\theta^{(2)}}E_{\theta_1}\varphi|T^{(2)} \\
&= E_\theta\varphi, \quad \forall\theta \in \Theta_{H_A}.
\end{aligned}$$

So φ^0 is UMP within $\mathcal{C}_S$. To prove (i), it is to be noted that

$$E_{\theta_1}\left(\varphi^0|T^{(2)} = t^{(2)}\right) \uparrow \theta_1$$

$$\Rightarrow E_{\theta_1}\left(\varphi^0 | T^{(2)} = t^{(2)}\right) \leq E_{\theta_{10}}\left(\varphi^0 | T^{(2)} = t^{(2)}\right) = \alpha, \forall \theta_1 \leq \theta_{10}$$

$$\Rightarrow E_\theta \varphi^0 = E_{\theta_1}\left(\varphi^0 | T^{(2)} = t^{(2)}\right) \leq \alpha, \forall \theta \in \Theta_H$$

$$\Rightarrow \varphi^0 \in \mathcal{C}_U.$$

Example 1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ independently where μ, σ^2 are both unkown.

$$p(x, \mu, \sigma^2) = (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tag{1}$$

Consider the following hypothesiss,

$$H_1 : \mu \leq 0 \text{ vs. } H_{1A} : \mu > 0$$

We can write eqn 3 as,

$$\begin{aligned} p(x|\mu, \sigma^2) &= (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu}{\sigma^2}\right\} \\ &\equiv c(\theta_1, \theta^{(2)}) \exp\left\{\theta_1 T_1(x) + \theta^{(2)} T^{(2)}(x)\right\} h(x) \end{aligned}$$

where, $\theta_1 = \frac{n\mu}{\sigma^2}$, $\theta^{(2)} = -\frac{1}{2\sigma^2}$, $T_1(x) = \bar{\mathbf{x}}$, and $T^{(2)}(x) = \sum_{i=1}^{n} x_i^2$. Hence H_1 can be rewritten as, $H_1 : \theta_1 = 0$. Then, $\Theta_B = \left\{(\theta_1, \theta^{(2)}) : \theta_1 = 0\right\}$. Note that, if $(T_1, T^{(2)})$ are jointly sufficient for $(\theta_1, \theta^{(2)}) \in \Theta$, then T is sufficient for Θ_B, which is also complete and hence boundedly complete, thus the test given by

$$\begin{aligned} \varphi(t_1, t^{(2)}) &= 1 \text{ if, } t_1 \geq c(t^{(2)}) \\ &= 0 \text{ if, } t_1 < c(t^{(2)}) \end{aligned} \tag{2}$$

with

$$E\left\{\varphi(T_1, T^{(2)}) | T^{(2)} = t^{(2)}\right\} = \alpha \text{ a.e. } \mathcal{P}_B^T \tag{3}$$

is uniformly most powerful (UMP) $\alpha-$similar test. Define,

$$V(T_1, T^{(2)}) = \frac{T_1}{\sqrt{T^{(2)} - nT_1^2}}, \quad \uparrow T_1 | T^{(2)}.$$

Note that,

$$
\begin{aligned}
V(T_1, T^{(2)}) &= \frac{\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \\
&= \frac{1}{\sqrt{n(n-1)}} \frac{\sqrt{n}\bar{X}/\sigma^2}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\sigma^2}} \\
&\sim t_{n-1} \text{ under } \Theta_B.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& P_{\theta_1=0}\left[T_1 > c(T^{(2)}) | T^{(2)} = t^{(2)}\right] \\
&= P_{\theta_1=0}\left[V(T_1, T^{(2)}) > c(t^{(2)}) | T^{(2)} = t^{(2)}\right] \\
&= P_{\theta_1=0}\left[V(T_1, T^{(2)})\sqrt{n(n-1)} > t_{\alpha;n-1}\right] \\
&= \alpha.
\end{aligned}
$$

Hence the test,

$$
\begin{aligned}
\varphi(t_1, t^{(2)}) &= 1 \text{ if, } \frac{\sqrt{n}\bar{X}}{s} > t_{\alpha;n-1} \\
&= 0 \text{ if, } \frac{\sqrt{n}\bar{X}}{s} \leq t_{\alpha;n-1}
\end{aligned}
$$

is UMPU size α test.

In the above example we see that to evaluate the cut-off points of the test we need the conditional distribution of $T_1 | T^{(2)} = t^{(2)}$ which may not be easy to obtain in all cases and also it may be difficult to work with them. So, it will be better if those conditional tests can be converted into unconditional tests by some means and it is possible to do so if there exists a suitable function $V = V(T_1, T^{(2)})$of the sufficient statistics $(T_1, T^{(2)})$ which is independent of $T^{(2)}$. Our aim is to find a statistic V such that

(i) V is strictly increasing function of T_1 for fixed $T^{(2)}$.

(ii)V and $T^{(2)}$ are independent.

Then the conditionally best test is given by

$$
\varphi^0(v, t^{(2)}) = 1 \text{ if } v > v(t^{(2)})
$$

$$= \quad \gamma(t^{(2)}) \text{ if } v = v(t^{(2)})$$
$$= \quad 0 \text{ if } v < v(t^{(2)})$$

where $v(t^{(2)})$ and $\gamma(t^{(2)})$ are such that $E_{\theta_{10}}\left[\varphi^0(V,T^{(2)})|T^{(2)} = t^{(2)}\right] = \alpha$.

Since V and $T^{(2)}$ are independently distributed the UMPU test is given by

$$\begin{aligned} \varphi^0(v) &= 1 \text{ if } v > v_0 \\ &= \gamma \text{ if } v = v_0 \\ &= 0 \text{ if } v < v_0 \end{aligned}$$

where v_0 and γ are such that $E_{\theta_{10}}[\varphi^0(V)] = \alpha$.

Choice of V

The statistic can be obtained by using the concept of ancillarity which has been already discussed in Module-12. We know that a statistics V is ancillary for $\mathcal{P}$ (or, for $\Theta^\star$) if the distribution of V is independent of any $P \in \mathcal{P}$ (or, any $\theta \in \Theta^\star$).

Basu's theorem

Suppose T is boundedly complete for $\Theta^\star$ and V is ancillary for $\Theta^\star$. Then for V and T are independent on $\Theta^\star$.

In the last example, the distribution of $V(T_1, T^{(2)})$ is independent of Θ_B and $T^{(2)}$ is complete sufficient for Θ_B. So $V(T_1, T^{(2)})$ and $T^{(2)}$ are independent.

Example 2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ independently where μ, σ^2 are both unknown.

$$p(x, \mu, \sigma^2) = (2\pi)^{-n/2}\,\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tag{4}$$

Consider the following hypothesis,

$$H_2 : \sigma^2 \leq \sigma_0^2 \text{ vs. } H_{2A} : \sigma^2 > \sigma_0^2$$

Similarly in this case, we can rewrite eqn 3

$$p(x|\mu, \sigma^2) \quad = \quad (2\pi)^{-n/2}\,\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu}{\sigma^2}\right\}$$

$$
\begin{aligned}
&= (2\pi)^{-n/2}\,\sigma^{-n}\exp\left\{-\left(\frac{1}{2\sigma^2}-\frac{1}{2\sigma_0^2}\right)\sum_{i=1}^{n}x_i^2+\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i-\frac{n\mu}{\sigma^2}\right\}\\
&\times \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}x_i^2\right\}\\
&\equiv c(\theta_1,\theta^{(2)})\exp\left\{\theta_1 T_1(x)+\theta^{(2)}T^{(2)}(x)\right\}+h(x)
\end{aligned}
$$

where $\theta_1 = -\frac{1}{2}\left(\frac{1}{\sigma^2}-\frac{1}{\sigma_0^2}\right)$, $\theta^{(2)} = \frac{\mu}{\sigma^2}$, $T_1(x) = \sum_{i=1}^{n}x_i^2$, $T^{(2)}(x) = \sum_{i=1}^{n}x_i$. Hence $\Theta_B = \left\{(\theta_1,\theta^{(2)}) : \theta_1 = 0\right\}$, and $H_2 \Leftrightarrow\ H_2 : \theta_1 \leq 0$ and $H_2 \Leftrightarrow\ H_{2A} : \theta_1 > 0$ UMP $\alpha-$similar test is given by,

$$
\begin{aligned}
\varphi(t_1,t^{(2)}) &= 1 \text{ if, } t_1 \geq c(t^{(2)})\\
&= 0 \text{ if, } t_1 < c(t^{(2)})
\end{aligned}
$$

with $E_{\theta_1=0}\varphi(T_1,T^{(2)}) = \alpha$. Under $\theta_1 = 0 \Leftrightarrow \sigma^2 = \sigma_0^2, T = \sum_{i=1}^{n}x_i$ is sufficient for θ. It is also complete. Here the corresponding ancillary statistics is, $V = \sum_{i=1}^{n}(X_i-\bar{X})^2$. [When $\theta_1 = 0$, $X_i - \bar{X}$ are all identically distributed and are independent of all nuisance parameters.] By Basu's theorem, V and T are independently distributed whenever $\theta_1 = 0$. Define,

$$
V = \sum_{i=1}^{n}(X_i-\bar{X})^2 = T_1 - \frac{T^{(2)}}{n} \uparrow T_1|T^{(2)} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}.
$$

Hence,

$$
\begin{aligned}
\varphi(t_1,t^{(2)}) &= 1 \text{ if, } \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha;n-1}\\
&= 0 \text{ if, } \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{\alpha;n-1}
\end{aligned}
$$

Example 3. Let $X_1 \sim B(n_1,\theta_1)$ and $X_2 \sim B(n_2,\theta_2)$ are independent. Consider the following testing problem for hypothesis,

$$
H : \theta_1 = \theta_2 \text{ vs. } H_A : \theta_1 > \theta_2.
$$

$$
\begin{aligned}
p(x|\theta_1,\theta_2) &= \binom{n_1}{x_1}\binom{n_2}{x_2}\theta_1^{x_1}(1-\theta)^{n_1-x_1}\theta_2^{x_2}(1-\theta)^{n_2-x_2}\\
&= \binom{n_1}{x_1}\binom{n_2}{x_2}\exp\left\{x_1\tau+(x_1+x_2)\theta\right\},
\end{aligned}
$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

$\tau = \left(\log \frac{\theta_1}{1-\theta_1} - \log \frac{\theta_2}{1-\theta_2}\right) = \log(\text{ odds ratio })$ and $\theta = \log \frac{\theta_2}{1-\theta_2}$. Define, $S(x) = X_1$ and $T(X) = X_1 + X_2$. An UMPS level α test is given by

$$\begin{aligned} \varphi(s,t) &= 1 \text{ if, } s > c(t) \\ &= a(t) \text{ if, } s = c(t) \\ &= 0 \text{ if, } s < c(t) \end{aligned}$$

where $c(t)$ and $a(t)$ are such that, $E_{\tau=0}\varphi(S,T) = \alpha$ for all t. Under $\tau = 0$, $S|T = t$ has the following p.m.f.,

$$P\left[S = s|T = t\right] = \frac{\binom{n_1}{s}\binom{n_2}{t-s}}{\binom{n_1+n_2}{t}}.$$

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
UMPU tests for multi-parameter exponential family-III

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

In the last module we have discussed the construction of UMPU test $(I) H : \theta_1 \leq \theta_{10}$ vs. $H_A : \theta_1 > \theta_{10}$,. In the present module we consider the UMPU test for testing $(II) H_0 : \theta_1 = \theta_{10}$ vs. $H_1 : \theta_1 \neq \theta_{10}$. Define, $g_\theta(\mathbf{t}) = c(\theta) \exp^{\sum_{i=1}^k \theta_i t_i} H(\mathbf{t})$. So, $T_1 | t^{(2)} \sim OPEF(\theta_1)$.

$$\begin{aligned}
\varphi^0(\mathbf{t}) &= 1 \text{ if, } t_1 < c_1(t^{(2)}) \text{ or, } t_1 > c_2(t^{(2)}) \\
&= \gamma_i(t^{(2)}) \text{ if, } t_1 = c_i(t^{(2)}), i = 1, 2 \\
&= 0 \text{ if, } c_1(t^{(2)}) < t_1 < c_2(t^{(2)}),
\end{aligned}$$

where $c_1, c_2, \gamma_1, \gamma_2$ are such that,

$$E_{\theta_{10}} \varphi^0 | t^{(2)} = \alpha, \tag{1}$$

$$E_{\theta_{10}} T_1 \varphi^0 | t^{(2)} = \alpha E_{\theta_{10}} T_1 | t^{(2)}, \tag{2}$$

φ^0 is conditionally UMP (given $t^{(2)}$) in $CU(\alpha, \theta_{10}) = \{\varphi : (1)$ and, (2) are satisfied. $\}$

To show that φ^0 is UMPU (unconditionally) at level α. Let φ be any unbiased test for testing

$$H_0 : \theta_1 = \theta_{10} \text{ vs. } H_1 : \theta_1 \neq \theta_{10}$$

$$\varphi \text{ is unbiased } \Rightarrow \varphi \text{ is } \alpha- \text{ similar on } \Theta_B \text{ (by result 1) }.$$

$$\Rightarrow \varphi \text{ has a Neymar Structure with respect } T^{(2)} \text{ over } \Theta_B. \text{ (by result 3) }.$$

$$\Rightarrow E_{\theta_{10}} \varphi | t^{(2)} = \alpha \quad \text{a.e.}$$

φ is unbiased [*i.e.*, $E_\theta\varphi$ has a minimum at $\theta = \theta_0$.]

$$\Rightarrow \frac{\partial E_\theta \varphi}{\partial \theta_{10}} = 0.$$

Now,

$$\begin{aligned}
E_\theta\varphi &= \int \varphi c(\theta) \exp^{\theta_1 t_1 + \sum_{i=2}^{k} \theta_i t_i} H(t) dt \\
\frac{\partial E_\theta \varphi}{\partial \theta_{10}} &= E_{\theta_{10}} T_1 \theta + \frac{\frac{\partial c(\theta)}{\partial \theta_{10}}}{c(\theta_{10})} E_{\theta_{10}} \varphi = 0 \\
or \quad & E_{\theta_{10}} T_1 \theta - E_{\theta_{10}} T_1 E_{\theta_{10}} \varphi = 0. \qquad (3)
\end{aligned}$$

Therefore, $\frac{\partial E_\theta \varphi}{\partial \theta_{10}} = 0$,

$$\begin{aligned}
&\Leftrightarrow \left[\int t_1 \varphi c(\theta) \exp^{\theta_1 t_1 + \sum_{i=2}^{k} \theta_i t_i} H(t) dt + \frac{\frac{\partial c(\theta)}{\partial \theta_{10}}}{c(\theta_{10})} \int \varphi c(\theta) \exp^{\sum_{i=1}^{k} \theta_i t_i} H(t) dt \right] \Bigg|_{\theta_1 = \theta_{10}} = 0 \\
&\Leftrightarrow E_{\theta_{10}} T_1 \varphi + \left(\frac{\frac{\partial c(\theta)}{\partial \theta_{10}}}{c(\theta_{10})} \Bigg|_{\theta_1 = \theta_{10}} \right) \alpha = 0 \\
&\Leftrightarrow E_{\theta_{10}} T_1 \varphi - \alpha E_{\theta_{10}} T_1 = 0 \\
&\Leftrightarrow E_{\theta_{10}} \left[E_{T_1 | T^{(2)}} \left(T_1 - E_{\theta_{10}} T_1 \right) \varphi \Big| \left(T^{(2)} = t^{(2)} \right) \right] = 0 \\
&\Leftrightarrow E_{T_1 | T^{(2)}} \left(T_1 - E_{\theta_{10}} T_1 \right) \varphi \Big| t^{(2)} = 0 \\
&\Leftrightarrow E_{\theta_{10}} \left(T_1 \varphi \Big| t^{(2)} \right) = \alpha E_{\theta_{10}} \left(T_1 \Big| t^{(2)} \right) \qquad (4)
\end{aligned}$$

Hence, $\varphi \in CU(\alpha, \theta_{10})$. So,

$$E_\theta \varphi^0 | t^{(2)} \geq E_\theta \varphi | t^{(2)} \ \forall \ t^{(2)}, \theta \text{ such that } \theta_1 \neq \theta_{10}$$

or,

$$E_\theta \varphi^0 \geq E_\theta \varphi \ \forall \ \theta \text{ such that } \theta_1 \neq \theta_{10}.$$

But φ is any unbiased level α test. Hence φ is unconditionally UMPU at level α for testing,

$$H_0 : \theta_1 = \theta_{10} \text{ vs. } H_1 : \theta_1 \neq \theta_{10}$$

Alternatve respresentation.

Find a statistic $V = f(T_1, T^{(2)})$ such that,

(i) V and $T^{(2)}$ are i.i.d. under $\Theta_B = \{\theta; \theta_1 = \theta_{10}\}$

(ii) $V = a(t^{(2)}) + T_1 b(t^{(2)})$ for each fixed $t^{(2)} \& b(t^{(2)}) > 0 \forall t^{(2)}$.

$$
\begin{aligned}
& E_{\theta_{10}} T_1 \varphi^0 | t^{(2)} = \alpha E_{\theta_{10}} T_1 | t^{(2)} \\
& \Leftrightarrow E_{\theta_{10}} \left(\frac{V - a(t^{(2)})}{b(t^{(2)})} \right) \varphi^0 | t^{(2)} = \alpha E_{\theta_{10}} \left(\frac{V - a(t^{(2)})}{b(t^{(2)})} \right) | t^{(2)} \\
& \Leftrightarrow E_{\theta_{10}} V \varphi^0 (V, t^{(2)}) | t^{(2)} = \alpha E_{\theta_{10}} V | t^{(2)}.
\end{aligned}
$$

Therefore,

$$t_1 < c_1(t^{(2)}) \Leftrightarrow v < c_1^\star(t^{(2)})$$

and,

$$t_1 > c_2(t^{(2)}) \Leftrightarrow v > c_2^\star(t^{(2)})$$

Hence the alternative UMPU test is given as,

$$
\begin{aligned}
\varphi^0(v) &= 1 \text{ if, } v < c_1 \text{ or, } v > c_2 \\
&= \gamma_i \text{ if, } v = c_i, i = 1, 2 \\
&= 0 \text{ if, } c_1 < v < c_2,
\end{aligned}
$$

where, $c_1, c_2, \gamma_1, \gamma_2$ are such that,

$$E_{\theta_{10}} \varphi^0(V) = \alpha$$

and,

$$E_{\theta_{10}} V \varphi^0(V) = \alpha E_{\theta_{10}} V.$$

Choice of V

(i) V is ancillary for Θ_B

(ii) $V = a(t^{(2)}) + T_1 b(t^{(2)})$ i.e. V is linear function of T_1.

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

Note:

If V is symmetric about a under Θ_B the UMPU test can be given by,

$$\varphi^0(v) = 1 \text{ if, } |v-a| > c$$
$$\gamma \text{ if, } |v-a| = c$$
$$0 \text{ if, } |v-a| < c,$$

where, $E_{\theta_{10}}\varphi^0(V) = \alpha$.

Example:

Ex 1. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. We are interested in testing,

$$(i) H_0 : \sigma = \sigma_0 \text{ vs. } H_1 : \sigma \neq \sigma_0$$

and,

$$(ii) H_0 : \mu = 0 \text{ vs. } H_1 : \mu \neq 0$$

Solution (i)

$$p_\theta(\mathbf{x}) = (2\pi)^{-n/2}\sigma^{-n}\exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{n\bar{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}}.$$

Let us choose,

$$\theta_1 = -\frac{1}{2\sigma^2} \text{ and, } T_1 = \sum_{i=1}^n X_i^2$$

and,

$$\theta_2 = \frac{\mu}{\sigma^2} \text{ and, } T^{(2)} = n\bar{X}$$

Define,

$$V = T_1 - \frac{\left(T^(2)\right)^2}{n}$$

where $a(t^{(2)}) = -\frac{T_2^2}{n}$ and $b(t^{(2)}) = 1$, a linear function of T_1 for each fixed $t^{(2)}$ also V and $T^{(2)}$ are independent. Then the UMPU test is given by,

$$\varphi^0(v) = 1 \text{ if, } \frac{v}{\sigma_0^2} < c_1 \text{ or, } \frac{v}{\sigma_0^2} > c_2$$

$$0 \text{ if, } c_1 < \frac{v}{\sigma_0^2} < c_2,$$

where, c_1 and c_2 are such that,

$$E_{\sigma_0^2}\varphi^0(V) = \alpha$$

and,

$$E_{\sigma_0^2}V\varphi^0(V) = \alpha E_{\sigma_0^2}V$$

Thus,

$$P\left[\frac{c_1}{\sigma_0^2} < \chi^2_{n-1} < \frac{c_2}{\sigma_0^2}\right] = 1 - \alpha$$

and,

$$P\left[\frac{c_1}{\sigma_0^2} < \chi^2_{n+1} < \frac{c_2}{\sigma_0^2}\right] = 1 - \alpha$$

Solution (ii) Now let us now re-define,

$$\theta_1 = \frac{\mu}{\sigma^2} \text{ and, } T_1 = n\bar{X}$$

and,

$$\theta_2 = -\frac{1}{2\sigma^2} \text{ and, } T^{(2)} = \sum_{i=1}^{n} X_i^2.$$

Define,

$$V = \frac{\frac{T_1}{\sqrt{n}}}{\sqrt{\left(T^{(2)} - \frac{T_1^2}{n}\right)/(n-1)}}$$

is an ancillary for $\Theta_B = \{\theta; \mu = 0\}$. Define, $W = \frac{T_1/\sqrt{n}}{\sqrt{T^{(2)}}}$ is ancillary for Θ_B since it is independent of σ and linear in T_1 and W is symmetric about 0. The UMPU test is given by,

$$\varphi^0(w) = 1 \text{ if, } |w| > c$$

$$0 \text{ if, } |w| < c,$$

where, $E_{\mu_0}\varphi^0(W) = \alpha.$

Ex 2. If $(X,Y) \sim N_2(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ and let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an i.i.d sample from (X,Y). We are interested in finding an UMPU test for,

$$(i) H_0 : \rho \le 0 \text{ vs. } H_1 : \rho > 0$$

and,

$$(ii) H_0 : \rho = 0 \text{ vs. } H_1 : \rho \ne 0$$

Solution Let $\theta = (\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ and $\Theta_B = \{\theta;\rho = 0\}$. Now the joint density is given as,

$$\begin{aligned} p_\theta(\mathbf{x},\mathbf{y}) &= (2\pi)^{-n/2}(\sigma_1\sigma_2\sqrt{1-\rho^2})^{-n} \\ &\times \exp^{-\frac{1}{(1-\rho^2)}\left[\sum_{i=1}^n (\frac{x_i-\mu_1}{\sigma_1})^2 + \sum_{i=1}^n (\frac{y_i-\mu_2}{\sigma_2})^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^n (x_i-\mu_2)(y_i-\mu_2)\right]} \\ &= f(\theta)\exp^{\frac{\rho}{\sigma_1\sigma_2(1-\rho^2)}\sum_{i=1}^n x_i y_i - \frac{1}{\sigma_1^2(1-\rho^2)}\sum_{i=1}^n x_i^2 - \frac{1}{\sigma_2^2(1-\rho^2)}\sum_{i=1}^n y_i^2}. \end{aligned}$$

So let us now define, $\theta_1 = \frac{\rho}{\sigma_1\sigma_2(1-\rho^2)}$ and $T_1 = \sum_{i=1}^n X_i Y_i$. Also $T^{(2)} = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i^2, \bar{X}, \bar{Y})$. So accordingly, we can define,

$$V = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2 \sum_{i=1}^n (Y_i-\bar{Y})^2}} = \frac{T_1 - g_1(T^{(2)})}{g_2(T^{(2)})} \uparrow t_1,$$

and linear in t_1 for each fixed $t^{(2)}$. Hence UMPU test for,

$$H_0 : \rho \le 0 \text{ vs. } H_1 : \rho > 0$$

is given by,

$$\begin{aligned} \varphi^0(v) = 1 &\text{ if, } \frac{\sqrt{n-2}v}{\sqrt{1-v^2}} > t_{\alpha;n-2} \\ 0 &\text{ if, } \frac{\sqrt{n-2}v}{\sqrt{1-v^2}} < t_{\alpha;n-2}, \end{aligned}$$

and the UMPU test for,

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \ne 0$$

A Gateway to All Post Graduate Courses

is given by,

$$\varphi^0(v) = 1 \text{ if, } \frac{\sqrt{n-2}|v|}{\sqrt{1-v^2}} > t_{\alpha/2;n-2}$$
$$0 \text{ if, } \frac{\sqrt{n-2}|v|}{\sqrt{1-v^2}} < t_{\alpha/2;n-2},$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Statistical Inference I
Theory of Confidence Sets

Shirsendu Mukherjee

Department of Statistics, Asutosh College, Kolkata, India.

shirsendu_st@yahoo.co.in

Introduction

The idea of interval estimation can be extended to include simultaneous estimation of several parameters. For example, two parameters (μ, σ^2) of a $N(\mu, \sigma^2)$ may be estimated by some subset S of the parameter space $\Theta = \mathcal{R} \times \mathcal{R}^+$. A $(1-\alpha)\%$ confidence region is a subset Θ such that if samples were repeatedly drawn from the population and a region constructed for each sample then about $(1-\alpha)\%$ of those regions would include true parameter value. So our discussion can be made somewhat more general if we speak in terms of confidence sets rather than confidence intervals.

Confidence set estimation

Let θ be the parameter of interest and Θ is the parametric space. While X is the random vector and χ is the support. Let also P_θ be the probability distribution of X.

Definition.(Confidence set)

A family of subsets $S(\mathbf{x})$ of Θ is said to constitute a family of confidence sets for θ at confidence level $(1-\alpha)$ if,

$$P_\theta\left[\theta \in S(x)\right] \geq 1-\alpha$$

Special cases:

I. **Lower confidence interval.** Lower confidence bound provides just a lower bound for the parameter of interest.

$$P_\theta\left[\theta \geq \underline{\theta}(x)\right] \geq 1-\alpha, \quad \forall \theta \in \Theta$$

Confidence set $= [\underline{\theta}(x), \infty)$ and $S(x) = \{\theta : \underline{\theta}(x) \leq \theta\}$.

epathshala A Gateway to All Post Graduate Courses

Example 1. Let $X_1, \ldots, X_n \sim N(\mu, 1)$. Since we know, $\bar{X} \sim N(\mu, \frac{1}{\sqrt{n}})$. Therefore,

$$P_\mu \left[\sqrt{n}\left(\bar{X} - \mu\right) \leq \tau_\alpha\right] = 1 - \alpha \ \forall \mu \in R.$$

Hence by interchanging sides,

$$P_\mu \left[\mu \geq \bar{X} - \frac{\tau_\alpha}{\sqrt{n}}\right] = 1 - \alpha \ \forall \mu \in R.$$

II. **Upper confidence interval.** Upper confidence bound provides just a upper bound for the parameter of interest.

$$P_\theta \left[\theta \leq \bar{\theta}(x)\right] \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

Confidence set $= \left(-\infty, \bar{\theta}(x)\right]$ and $S(x) = \left\{\theta : \bar{\theta}(x) \geq \theta\right\}$.

III. **Confidence interval.** Confidence interval provides both upper and lower bound for the parameter of interest.

$$P_\theta \left[\underline{\theta}(x) \leq \theta \leq \bar{\theta}(x)\right] \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

Confidence set $= \left[\underline{\theta}(x), \bar{\theta}(x)\right]$ and $S(x) = \left\{\theta : \bar{\theta}(x) \leq \theta \leq \bar{\theta}(x)\right\}$.

Relation between confidence set estimation and testing of hypothesis.

There is a very strong correspondence between hypothesis testing and interval estimation. In fact we can say in general that every confidence set corresponds to a test and vice-versa. We next consider the method of test inversion and explore the relationship between a test of hypothesis for a parameter θ and confidence interval for θ.

Result : Let for each $\theta_0 \in \Theta$, $A(\theta_0)\,(\subset \mathcal{X})$ be the acceptance region of a level α test for testing $H_0(\theta) : \theta = \theta_0$. Then,

$$S(x) = \{\theta : \mathbf{x} \in \mathbf{A}(\theta)\}$$

epathshala A Gateway to All Post Graduate Courses

is a confidence set for θ, at confidence level $1-\alpha$.

Proof. Let $R(\theta_0)$ be the rejection region of the level α test mentioned above for testing $H(\theta_0) : \theta = \theta_0$. We know that $A(\theta_0) \cup R(\theta_0) = \mathcal{X}$ (sample space).

$$\begin{aligned} P_{\theta_0}\left[X \in R(\theta_0)\right] &\leq \alpha \quad \forall \theta_0 \in \Theta \\ \text{or, } P_{\theta}\left[X \in R(\theta)\right] &\leq \alpha \quad \forall \theta \in \Theta \\ \text{or, } P_{\theta}\left[X \in A(\theta)\right] &\geq 1-\alpha \quad \forall \theta \in \Theta \\ \text{or, } P_{\theta}\left[X \in S(X)\right] &\geq 1-\alpha \quad \forall \theta \in \Theta \end{aligned}$$

Example Let $\mathbf{X} = X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Suppose we are interested in the following testing hypothesis,

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

We know that, we reject H_0 at level α if $\frac{\sqrt{n}\left|\bar{X}-\mu_0\right|}{s} > t_{\alpha/2,n-1}$. Therefore,

$$R(\mu_0) = \left\{x : \frac{\sqrt{n}\left|\bar{X}-\mu_0\right|}{s} > t_{\alpha/2,n-1}\right\}.$$

$$R(\mu) = \left\{x : \frac{\sqrt{n}\left|\bar{X}-\mu\right|}{s} > t_{\alpha/2,n-1}\right\}.$$

$$A(\mu) = \left\{x : \frac{\sqrt{n}\left|\bar{X}-\mu\right|}{s} \leq t_{\alpha/2,n-1}\right\} = \left\{x : \mu \in S(x)\right\},$$

where $S(x) = \{\mu : x \in A(\mu)\}$ Hence,

$$A(\mu) = \left\{x : \bar{x} - \frac{x}{\sqrt{n}} t_{\alpha/2,n-1} \leq \mu \leq \bar{x} + \frac{x}{\sqrt{n}} t_{\alpha/2,n-1}\right\},$$

and

$$S(x) = \left[\bar{x} - \frac{x}{\sqrt{n}} t_{\alpha/2,n-1}, \bar{x} + \frac{x}{\sqrt{n}} t_{\alpha/2,n-1}\right].$$

Pathshala पाठशाला A Gateway to All Post Graduate Courses

Confidence interval for discrete case

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be random sample of size n drawn from a discrete distribution with p.m.f. $p(x, \theta)$. To determine a $100(1-\alpha)\%$ confidence interval for θ we may use the distribution of the sufficient statistic T for θ. We consider the following result without proving it. For a proof of the result one may see Casella and Berger.

Result Let T be a discrete statistic with c.d.f. $F_\theta(t) = P_\theta(T \leq t)$. Let α_1 and α_2 be two fixed real numbers such that $\alpha_1 + \alpha_2 = \alpha$. Suppose that for each t, $\underline{\theta}(t)$ and $\bar{\theta}(t)$ is defined as follows:

(i) If $F_\theta(t)$ is a decreasing function of θ for each t, define $\underline{\theta}(t)$ and $\bar{\theta}(t)$ by

$$P_{\bar{\theta}(t)}(T \leq t) = \alpha_1 \text{ and } P_{\underline{\theta}(t)}(T \geq t) = \alpha_2.$$

(ii) If $F_\theta(t)$ is a increasing function of θ for each t, define $\underline{\theta}(t)$ and $\bar{\theta}(t)$ by

$$P_{\bar{\theta}(t)}(T \geq t) = \alpha_1 \text{ and } P_{\underline{\theta}(t)}(T \leq t) = \alpha_2.$$

Then the interval $[\underline{\theta}(t), \bar{\theta}(t)]$ is a confidence interval for θ with confidence coefficient $(1-\alpha)$.

Example Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $Poisson(\theta)$ distribution. Here $T = \sum_{i=1}^n X_i$ is sufficient for θ. Since $T \sim Poisson(n\theta)$, the c.d.f. of T is a decreasing function of θ. If $T = t$ is observed, taking $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, we have to solve $\underline{\theta}(t)$ and $\bar{\theta}(t)$ from the equations

$$\sum_{i=0}^{t} \frac{e^{-n\bar{\theta}(t)}(n\bar{\theta}(t))^t}{t!} = \frac{\alpha}{2} \text{ and } \sum_{i=t}^{\infty} \frac{e^{-n\underline{\theta}(t)}(n\underline{\theta}(t))^t}{t!} = \frac{\alpha}{2}$$

Now we know that if $X \sim Poisson(\lambda)$ then for any non-negative integer k

$$P(X \leq k) = P(Y \geq 2\lambda), \text{ where } Y \sim \chi^2_{2(k+1)}$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

Hence

$$\sum_{i=0}^{t} \frac{e^{-n\bar{\theta}(t)}(n\bar{\theta}(t))^t}{t!} = \frac{\alpha}{2}$$

$$\Rightarrow P(\chi^2_{2(t+1)} \geq 2n\bar{\theta}(t)) = \frac{\alpha}{2}$$

$$\Rightarrow 2n\bar{\theta}(t) = \chi^2_{2(t+1),\frac{\alpha}{2}}$$

$$\Rightarrow \bar{\theta}(t) = \frac{1}{2n}\chi^2_{2(t+1),\frac{\alpha}{2}}.$$

Again

$$\sum_{i=t}^{\infty} \frac{e^{-n\underline{\theta}(t)}(n\underline{\theta}(t)^t}{t!} = \frac{\alpha}{2}$$

$$\Rightarrow P(\chi^2_{2t} \leq 2n\underline{\theta}(t)) = \frac{\alpha}{2}$$

$$\Rightarrow 2n\underline{\theta}(t) = \chi^2_{2t,1-\frac{\alpha}{2}}$$

$$\Rightarrow \underline{\theta}(t) = \frac{1}{2n}\chi^2_{2t,1-\frac{\alpha}{2}}.$$

Hence $[\frac{1}{2n}\chi^2_{2t,1-\frac{\alpha}{2}}, \frac{1}{2n}\chi^2_{2(t+1),\frac{\alpha}{2}}]$ is confidence interval for θ with confidence coefficient $(1-\alpha)$.

Optimum property of confidence sets

Since there is a one-to-one correspondence between confidence sets and test of hypotheses, there is some correspondence between optimality of tests and optimality of confidence sets. Usually test related optimality properties of confidence sets do not directly relate to the size of the set but rather to the probability of the set covering false values. The probability of covering false values or the probability of false coverage, indirectly measures the size of the confidence set. Intuitively, smaller sets cover few values and,hence, are less likely to cover false values. An optimum property of confidence sets is given in the following definition.

Uniformly most accurate confidence set (UMA)

A family of confidence sets $S^0(x), x \in \mathcal{X}$ is said to constitute a family of UMA

confidence sets at confidence level $(1-\alpha)$ if

$$(i) P_\theta\left(\theta \in S^0(X)\right) \geq 1-\alpha \quad \forall \theta \in \Theta$$
$$(ii) P_\theta\left(\theta' \in S^0(X)\right) \leq P_\theta\left(\theta' \in S(X)\right) \quad \forall \theta \neq \theta'$$

and whatever $S(X)$ satisfying (i).

A $1-\alpha$ confidence set that minimizes the probability of false coverage over a class of $1-\alpha$ confidence sets is called a UMA set. UMA confidence sets are constructed by inverting the acceptance region of UMP test,as we will prove below.

Result: Let for each $\theta_0 \in \Theta$, $A^0(\theta_0)$ be the acceptance region for a UMP level α test for testing $H_0(\theta_0) : \theta = \theta_0$. Then,

$$S^0(x) = \left\{\theta; x \in A^0(\theta)\right\}$$

is UMA at confidence level $(1-\alpha)$.

Proof. Let $R^0(\theta_0)$ be the rejection region for the UMP test mentioned above for every $\theta_0 \in \Theta$. By the previous result,

$$\begin{aligned}
P_\theta(\theta \in S(X)) &\geq 1-\alpha \quad \forall \theta \in \Theta \\
P_\theta\left[X \in R^0(\theta_0)\right] &\geq P_\theta\left[X \in R(\theta_0)\right] \quad \forall \theta \neq \theta_0 \\
P_\theta\left[X \in A^0(\theta_0)\right] &\geq P_\theta\left[X \in A(\theta_0)\right] \quad \forall \theta \neq \theta_0 \\
P_\theta\left[\theta' \in S^0(X)\right] &\geq P_\theta\left[\theta \in S(X)\right] \quad \forall \theta \neq \theta'.
\end{aligned}$$

and whatever $S(x)$ satisfying $P_\theta(\theta \in S(X)) \geq 1-\alpha$.

Example 1. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\theta, 1)$. Rejection region for a UMP test of size α for testing

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta > \theta_0.$$

is $\sqrt{n}\left(\bar{X}-\theta_0\right) > \tau_\alpha$ where τ_α is the upper α point of standard normal distribution, and define the acceptance region,

$$A(\theta_0) = \left\{\mathbf{x}; \bar{x} \leq \theta_0 + \frac{\tau_\alpha}{\sqrt{n}}\right\}$$

and,

$$S(x) = \left\{\mathbf{x}; \theta \geq \bar{x} - \frac{\tau_\alpha}{\sqrt{n}}\right\}.$$

Hence the UMA confidence interval is given,

$$\left[\bar{x} - \frac{\tau_\alpha}{\sqrt{n}}, \infty\right).$$

ePathshala पाठशाला A Gateway to All Post Graduate Courses

Note The more common two-sided interval $\left[\bar{x} - \frac{\tau_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{x} + \frac{\tau_{\frac{\alpha}{2}}}{\sqrt{n}}\right]$. is not UMA since it is obtained by inverting the acceptance region of the from the test of $H : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$ for which no UMP test exists.

Example 2. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim R(0, \theta)$. UMP test for testing

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

So we reject $H_0(\theta_0)$ if $x_{(n)} > \theta_0$ or, $x_{(n)} < \theta_0 \alpha^{1/n}$, and define the acceptance region,

$$A(\theta_0) = \left\{\mathbf{x}; \theta_0 \alpha^{1/n} \leq x_{(n)} \leq \theta_0\right\}$$

and,

$$S(x) = \left\{\theta; x_{(n)} \leq \theta_0 \leq \frac{x_{(n)}}{\alpha^{1/n}}\right\}$$

Hence the UMA confidence interval is given,

$$\left[x_{(n)}, \frac{x_{(n)}}{\alpha^{1/n}}\right].$$

Statistical Inference I
Theory of Unbiased Confidence Sets

Shirsendu Mukherjee
Department of Statistics, Asutosh College, Kolkata, India.
shirsendu_st@yahoo.co.in

In the last lecture we have studied test inversion as one of the methods of constructing confidence intervals. We showed that UMP tests lead to UMA confidence intervals. In Module 32 we have seen that UMP tests generally do not exist for two-sided alternatives. In such situations we either restrict consideration to smaller subclasses of tests by requiring that the test functions have some desirable properties, or we restrict the class of alternatives to those near the null parameter values. The construction of unbiased tests and locally most powerful tests were already discussed. In this lecture we discuss how the concept of unbiasedness can be used in constructing confidence intervals.

A Gateway to All Post Graduate Courses

Unbiased confidence set

A family of confidence sets $S(x), x \in \mathcal{X}$ for a parameter θ is said be unbiased at confidence level $(1-\alpha)$ if

$$(i) P_\theta\left(\theta \in S(X)\right) \geq 1-\alpha \;\; \forall \theta \in \Theta$$
$$(i) P_\theta\left(\theta' \in S(X)\right) \leq 1-\alpha \;\; \forall \theta \neq \theta'$$

If $S(x)$ is a is an interval satisfying (i) and (ii), we call it a $(1-\alpha)$-level unbiased confidence interval.

Unifromly most accurate unbiased (UMAU) confidence set:

A family of confidence sets $S^0(x), x \in \mathcal{X}$ for a parameter θ is said to constitute a family of UMAU confidence sets of confidence level $(1-\alpha)$ if it is unbiased at

confidence level $(1-\alpha)$ and

$$P_\theta\left(\theta' \in S^0(X)\right) \leq P_\theta\left(\theta' \in S(X)\right) \quad \forall \theta \neq \theta' \in \Theta.$$

where $S(X)$ is any other unbiased confidence set at confidence level $(1-\alpha)$.

Remark A a family $S(X)$ of confidence sets for a parameter θ is unbiased at level $1-\alpha$ if the probability of true coverage is at least $1-\alpha$ and that of false coverage is at most $1-\alpha$. In other words, $S(X)$ traps a true parameter value more often than it does a false one.

Result 1: Let for each $\theta_0 \in \Theta$, $A^0(\theta_0)$ be the acceptance region for a UMPU level α test for testing $H(\theta_0) : \theta = \theta_0$ against $H_A(\theta_0) : \theta \neq \theta_0$Then,

$$S^0(x) = \left\{\theta; x \in A^0(\theta)\right\}$$

is UMAU at confidence level $(1-\alpha)$.

Proof. To see that $S^0(x)$ is unbiased, we note that since $A^0(\theta)$ is the acceptance region of an unbiased test,

$$P_\theta(\theta' \in S^0(X)) = P_\theta\left[X \in A^0(\theta')\right] \leq (1-\alpha).$$

We next show that $S^0(X)$ is UMA. Let $S(x)$ be any other unbiased (1α)-level family of confidence sets, and write $A(\theta) = \{x : \theta \in S(x)\}$. Then $P_\theta\left[X \in A(\theta')\right] = P_\theta(\theta' \in S(X)) \leq (1-\alpha)$, and it follows that $A(\theta)$ is the acceptance region of an unbiased size a test. Hence

$$\begin{aligned}
P_\theta(\theta' \in S(X)) &= P_\theta\left[X \in A(\theta')\right] \forall \theta \in \Theta \\
&\geq P_\theta\left[X \in A^0(\theta')\right] \quad \forall \theta \neq \theta_0 \\
&= P_\theta\left[\theta' \in S(X)\right] \quad \forall \theta \neq \theta_0.
\end{aligned}$$

The inequality follows since $A^0(\theta)$ is the acceptance region of a UMPU test. This completes the proof.

Example 1. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\theta, 1)$. Rejection region for a UMPU size α test for testing

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

is $\sqrt{n}\left|\bar{X} - \mu_0\right| > \tau_{\alpha/2}$ where τ_α is the upper α point of standard normal distribution, and define the acceptance region,

$$A(\theta_0) = \left\{\mathbf{x}; \theta_0 - \frac{\tau_{\alpha/2}}{\sqrt{n}} \leq \bar{x} \leq \theta_0 + \frac{\tau_{\alpha/2}}{\sqrt{n}}\right\}$$

and,

$$S(x) = \left\{\mathbf{x}; \bar{x} - \frac{\tau_{\alpha/2}}{\sqrt{n}} \leq \theta \leq \bar{x} + \frac{\tau_{\alpha/2}}{\sqrt{n}}\right\}.$$

Hence the UMAU confidence interval is given,

$$\left[\bar{x} - \frac{\tau_{\alpha/2}}{\sqrt{n}}, \bar{x} + \frac{\tau_{\alpha/2}}{\sqrt{n}}\right].$$

Example 2. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$ where both μ and σ^2 are unknown.Rejection region for a UMPU size α test for testing

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

is $\sqrt{n}\left|\bar{X} - \mu_0\right| > st_{n-1,\alpha/2}$ where $t_{n-1,\alpha}$ is the upper α point of t-distribution with (n-1) d.f.and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and define the acceptance region,

$$A(\theta_0) = \left\{\mathbf{x}; \mu_0 - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \leq \bar{x} \leq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}\right\}$$

and,

$$S(x) = \left\{\mathbf{x}; \bar{x} - \frac{t_{\alpha/2}}{\sqrt{n}} \leq \theta \leq \bar{x} + \frac{t_{\alpha/2}}{\sqrt{n}}\right\}.$$

Hence the UMAU confidence interval is given,

$$\left[\bar{x} - \frac{t_{\alpha/2}}{\sqrt{n}}, \bar{x} + \frac{t_{\alpha/2}}{\sqrt{n}}\right].$$

Example 3. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\theta, \sigma^2)$. The UMPU test for testing,

$$H_0 : \sigma = \sigma_0 \text{ vs. } H_1 : \sigma \neq \sigma_0.$$

is given by,

$$\begin{aligned} \varphi(v) &= 1 \text{ if } \frac{v}{\sigma_0^2} < c_1 \text{ or } \frac{v}{\sigma_0^2} > c_2 \\ &= 0 \text{ otherwise,} \end{aligned}$$

where $v = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Therefore $\frac{V}{\sigma_0^2} \sim \chi^2_{n-1}$ under H_0. Hence, the acceptance region for the UMPU test is,

$$c_1 \leq \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} \leq c_2,$$

where c_1 and c_2 are obtained from the following conditions,

$$\begin{aligned} &(i) E_{\sigma_0^2}\varphi(V) = \alpha \\ &(ii) E_{\sigma_0^2} V\varphi(V) = \alpha E_{\sigma_0^2}\varphi(V). \end{aligned}$$

By inverting the acceptance region the UMAU confidence interval for σ^2 is given by

$$\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{c_2}, \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{c_1}\right].$$

Example 4. Let $\mathbf{X} = (X_1, \ldots, X_m) \sim N(\theta, \sigma^2)$, and $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim N(\mu, \sigma^2)$. Consider the parameter for difference of mean $\delta = \theta - \mu$ The accepting region for a UMPU test for testing,

$$H_0 : \delta = \delta_0 \text{ vs. } H_1 : \delta \neq \delta_0,$$

is

$$A(\theta_0) = \left\{ \mathbf{x}; \delta_0 - s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2} \leq \bar{x} \leq \delta_0 + s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2} \right\}$$

where $S^2 = \frac{\left[\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2\right]}{(m+n-2)}$ and,

$$S(x) = \left\{ \mathbf{x}; \bar{x} - s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2} \leq \delta \leq \bar{x} + s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2} \right\}.$$

Hence the UMAU confidence interval is given,

$$\left[\bar{x} - s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2}, \bar{x} + s\sqrt{\frac{1}{m} + \frac{1}{n}} t_{m+n-2,\alpha/2} \right].$$

epathshala पाठशाला A Gateway to All Post Graduate Courses

Relation between shortest length confidence set and UMA confident set: If the measure of precision of a confidence interval is its expected length, one is naturally led to a consideration of unbiased confidence intervals. Pratt (1961) has shown that the expected length of a confidence interval is the average of false coverage probabilities.

Result 2 Let Θ be an interval on the real line and $p(x, \theta)$ be the p.d.f. of X. Let $S(X)$ be a family of $(1 - \alpha)$-level confidence intervals of finite length. If the length of a confidence set $S(x)$ is $\lambda(S(x))$ = Lebesgue measure of set $S(x)$ then

$$E_\theta \left[\lambda(S(x))\right] = \int_{\theta' \neq \theta} P_\theta \left(\theta' \in S(x)\right) d\theta'.$$

Proof

$$\begin{aligned}
\lambda(S(x)) &= \int_{\theta' \in S(x)} d\theta' \\
\text{Therefore, } E_\theta \left[\lambda(S(x))\right] &= \int \lambda(S(x)) p(x, \theta) dx \\
&= \int \left[\int_{\theta' \in S(x)} d\theta' \right] p(x, \theta) dx \\
&= \int \left[\int_{\theta' \in S(x)} p(x, \theta) dx \right] d\theta', \text{ By Fubini's theorem}
\end{aligned}$$

$$
\begin{aligned}
&= \int P_\theta\left(\theta' \in S(x)\right) d\theta' \\
&= \int_{\theta' \neq \theta} P_\theta\left(\theta' \in S(x)\right) d\theta'.
\end{aligned}
$$

Remark If $S(X)$ be a UMA confidence interval then it is also the uniformly shortest expected length (USEL)confidence interval. Similarly UMAU confidence interval is equivalent to the uniformly shortest unbiased expected length (USUEL)confidence interval.

Randomized confidence set

The idea of inverting acceptance region to obtain the confidence set can not be directly used for discrete cases, since we have randomized test. In fact, in discrete problems inverting acceptance region of randomized tests may not lead to a confidence set with a given confidence coefficient. Note that randomization is used in hypothesis testing to obtain tests with a given size. Thus, the same idea can be applied to confidence sets, i.e., we may consider randomized confidence sets. Suppose a UMPU test for testing,

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0,$$

is given by,

$$
\begin{aligned}
\varphi_{\theta_0}(x) &= 1 \text{ if, } t_1 > c_2(t_2, \theta_0) \text{ or, } t_1 < c_1(t_2, \theta_0) \\
&= \gamma_1(x) \text{ if, } t_1 = c_1(x, \theta_0) \\
&= \gamma_2(x) \text{ if, } t_1 = c_2(t_2, \theta_0) \\
&= 0 \text{ if, } c_1(t_2, \theta_0) < t_1 < c_2(t_2, \theta_0).
\end{aligned}
$$

Let U be a random variable that is independent of X and has the uniform distribution $U(0, 1)$. Then the test $\tilde{\varphi}_{\theta_0}(X, U) = \mathcal{I}_{(U,1]}(\varphi_{\theta_0})$ has the same power as φ_{θ_0} and is non-randomized if U is viewed as a part of the sample. Let

$$A_U(\theta_0) = \{(x, U); U \geq \varphi_{\theta_0}(x)\}$$

be the acceptance region for $\widetilde{\varphi}_{\theta_0}(X,U)$. If $\varphi_{\theta_0}(x)$ has size α for all θ_0, then inverting $A_U(\theta_0)$ we obtain a confidence set,

$$C(X,U) = \{\theta; (x,U) \in A_U(\theta)\}.$$

Example 5. Let X and Y are independently distributed as the Poisson distribution $P(\lambda_1)$ and $P(\lambda_2)$ and suppose we are interested in finding a confidence interval of $\rho = \lambda_1/\lambda_2$. Define, $T_1 = X$ and $T_2 = X+Y$. We know that $T_1|T_2 = t_2 \sim Bin\left(t_2, \frac{\rho}{1+\rho}\right)$. A UMPU test for testing,

$$H_0 : \rho = \rho_0 \text{ vs. } H_1 : \rho \neq \rho_0,$$

is given by,

$$\begin{aligned}
\varphi_{\rho_0}(t_1,t_2) &= 1 \text{ if, } t_1 > c_2(t_2,\rho_0) \text{ or, } t_1 < c_1(t_2,\rho_0) \\
&= \gamma_1(t_2) \text{ if, } t_1 = c_1(t_2,\rho_0) \\
&= \gamma_2(t_2) \text{ if, } t_1 = c_2(t_2,\rho_0) \\
&= 0 \text{ if, } c_1(t_2,\rho_0) < t_1 < c_2(t_2,\rho_0).
\end{aligned}$$

The acceptance region for non-randomized test is,

$$A_U(\rho_0) = \{(t_1,t_2,U); U \geq \varphi_{\rho_0}(t_1,t_2)\},$$

and the confidence set,

$$C(T_1,T_2,U) = \{\theta; (T_1,T_2,U) \in A_U(\rho)\}.$$

Nonparametric Inference: Module 1[1]

What we provide in this module

- A genesis of nonparametric inference
- Parametric Procedures-The traditional practice with limitations
- What is Nonparametric statistics today?

  Nonparametric, the phrase

  Nonparametric & Distribution free procedures
- Advantages, Disadvantages & Recommendation
- Software based learning
- An exploratory example of real field

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A genesis of Nonparametric Inference

Although statistical methods are used from the ancient times, the discipline of nonparametric statistics is developed only in the earlier decades of the nineteenth century. Savage(1953) considered the year 1936 as the inception of the subject of nonparametric statistics for the publication of the work on rank correlation by Hotelling and Pabst (1936). However, in a review work, Scheffe (1943), indicated the presence of the sign test in Fishers seminal book length treatment *Statistical Methods for Research Workers* in 1925. Further works, bearing the development of nonparametric statistics, include contributions by Friedman (1937), Kendall (1938), and Smirnov (1939). But the most significant contribution in this context is due to Wilcoxon (1945), who developed the rank based methods for comparing unknown distributions. It was perhaps the earliest attempt to derive a statistical procedure, parallel to the two sample t test without any distribution assumption. This particular work played the key role in accelerating the development of rank-based statistical procedures in the 1950s and 1960s. In further works Pitman (1948), Hodges and Lehmann (1956), and Chernoff and Savage (1958) investigated the efficiency aspects of rank based tests and ended with the promising outcome relative to parametric competitors. These works popularized the use of nonparametric statistical procedures among research workers and attracted practitioners of real fields.

For a brief account of the developments up to the recent times in the field of nonparametric statistics, we refer the interested reader to a special issue of the journal *Statistical Science* which gives exposure of a wide variety of topics. These include articles on comparing variances and other dispersion measures by Boos and Brownie(2004), density estimation by Sheather(2004), quantile-quantile (QQ) plots by Marden(2004), spatial statistics by Chang(2004), reliability methods by Hollander and Pena(2004) and permutation tests by Ernst(2004), among others.

2 Parametric Procedures-The traditional practice with limitations

Most of the traditional statistical tools are based on the parametric assumption: the data at hand can be thought of as generated by some well-known distribution like normal, exponential or Poisson. The parameter(s) of the distributions are assumed unknown and in a parametric inference problem we try to put some idea about the unknown parameter in the form of estimation or testing or confidence interval estimation. However, most of the times the normal distribution is used as the underlying population. The assumption of normality is often justified by the Central Limit Theorem, which ensures closeness to normality of certain statistics for large enough sample sizes. Other distributions are also important in different fields of applications. For example lifetime of physical systems are often characterized by exponential, Weibull or Gamma distributions. In such cases the interest lies in knowing the expected failure times and the mentioned distributions are used to characterize the lifetime distribution. The lifetime distribution is also of interest in the field of certain medical trials, where the goal is to give an idea about the length of the life after certain treatments. Exponential, lognormal or Weibull are often used to model the underlying lifetime distribution so as to enable appropriate inferential procedures. Again, in the analysis of economic data, Pareto or lognormal are the appropriate to capture the features of the income distribution.

However, in practice, most of the experiments are of complex nature, affected by a number of factors and hence the generated data might not be identified by a well-known distributions. Although for large data sets, normality assumption can be made for analysis but often the amount of data is small and hence requires assumption of an appropriate distribution for analysis. In fact, there are no exploratory methods to find the appropriate underlying distribution or to differentiate among the available choices. Therefore, a traditional statistician has to assume a distribution for the data which captures only certain features of the data to find without caring for the quality of the inference.

3 What is Nonparametric statistics today ?

3.1 Nonparametric, the phrase

The specific word, *Nonparametric* has its root in the works of Jacob Wolfowitz (1942) , where he said

> *We shall refer to this situation where a distribution is completely determined by the knowledge of its finite parameter set as the parametric case and denote the opposite case, where the functional forms of the distributions are unknown as the non-parametric case.*

Therefore, parametric statistics is based on known distributions with unknown parameters and nonparametric statistics was defined in the opposite way. Randles, Hettmansperger and Casella (2004) echoed the same notion in the statement

> *Nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family.*

The limitations of traditional methods facilitate the development of further statistical techniques, which can be applied regardless of the true distribution of the data. These techniques are the well known nonparametric and distribution-free methods.

3.2 Nonparametric & Distribution free procedures

Although the terms nonparametric and distribution-free are invariably used, but these actually indicate statistical procedures, which are used in the absence of any assumption of the underlying distribution. In the words of Bradley(1968)

> *The terms nonparametric and distribution-free are not synonymous... Popular usage, however, has equated the terms ... Roughly speaking, a nonparametric*

test is one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.

We start with the definition of these two important concepts and identify the differences subsequently.

Distribution Free statistic: Let $X_i, i = 1, 2, , , , n$ be n variables with the joint distribution F, where $F \in \mathcal{F}$, a class of joint distributions. Then a statistic $T = T(X_1, .., X_n)$ is said to be distribution free over $\mathcal{F}$, if the distribution of T remains the same for every possible $F \in \mathcal{F}$.

For example, if F is the joint distribution of n iid normal variables with unknown mean μ and variance unity. Then the statistic $Z = \sum_{i=1}^{n}(X_i - \bar{X})^2$ has a χ^2 distribution with n-1 degrees of freedom and hence is distribution free over $\mathcal{F}$.

However, in the above example, the distribution free statistic is actually *ancillary* in parametric terminology, that is, they are independent of the indexing parameter of the joint distribution. But they are not nonparametric statistics, since their distributions vary for different joint distributions. For instance, the distribution of Z above is no longer χ^2 when the joint distribution is different from normal.

Nonparametric distribution Free statistic: A statistic $T = T(X_1, .., X_n)$ is nonparametric distribution free over $\mathcal{F}$, if the distribution of T does not depend on any $F \in \mathcal{F}$.

For example, if F is the joint distribution of n iid continuous random variables having symmetry at the origin, then the statistic $Z = \sum_{i=1}^{n} I(X_i > 0)$, has a binomial distribution with parameters n and success probability $\theta_F = P_F(X_1 > 0)$. Now due to symmetry at the origin, $\theta_F = \frac{1}{2}$ and hence the distribution of Z remains the same (i.e. $Binomial(n, \frac{1}{2})$) whatever the distribution F be. Thus Z is nonparametric distribution free.

However, for the sake of simplicity, we use the term *nonparametric* to refer statistical procedures based on a nonparametric distribution free statistic.

4 Advantages, Disadvantages & Recommendation

The main problem with parametric inference is that if the assumed distribution is not correct, then all the efforts might went into vein and the best (i.e. efficient) can become the "worst". Therefore, nonparametric procedures enjoy the following advantages

1. Nonparametric procedures are based on fewer assumptions about the underlying populations of the data. Therefore, these procedures can accommodate analysis of non normal data.

2. These techniques are easy to understand and are often easier to apply. Thus unlike parametric procedures, the calculations with the complex distributions can be avoided.

3. Although nonparametric procedures discard a portion of information about the data, but the loss in efficiency as compared to the competitors based on the normal parents is only nominal. However, such procedures tend to be more efficient when the underlying population deviates from normality.

4. Nonparametric methods are robust in the sense that it tolerates departures from assumptions. That is, if the distributional assumption is perfect, one can adopt the usual parametric procedures. But, if the assumption of distribution is not perfect, nonparametric methods are the best alternative.

But nonparametric methods are not always the most desirable procedures as.

1. Most of the nonparametric methods use only ranks or signs of the observations discarding the further features of the data. This makes such procedures less efficient.

2. Nonparametric methods are usually not as *efficient* as their parametric counterparts when the assumptions are met.

Although nonparametric methods require fewer assumptions but they are used rarely in practice because of the use of limited information contained in the data. Consequently, nonparametric methods are primarily recommended in situations when assumptions are grossly violated(e.g. when severe skewness in the data is present).

5 Software based learning

With the advent of high-speed computing facility and softwares like R, nonparametric techniques grow in a fast pace. We emphasise on using R for learning purpose. Actually R is an open source statistical software and users can obtain the software free of charge through the Comprehensive R Archive Network (CRAN). Most importantly, if the required statistical methodology is not available within R, one can also write codes for his/her purpose. Consequently, for data based problems, we not only give the final result but also provide the R commands or codes together with the output.

6 An exploratory example of real field

Consider the following problem:

> *Suppose the average weekly sales of a new cell phone are collected for 12 consecutive weeks from a particular shop are as below*
>
> 52 39 49 33 58 61 44 221 201 133 289 211
>
> *The average weekly sales was 125 units for the last year. Is the evidence sufficient to conclude that sales this year exceeds last years sales?*

The usual statistical procedure test in this context is a single sample t test. Specifically, if μ denotes the true weekly sales, then we are interested in testing $H_0 : \mu = 125$ against $H_1 : \mu > 125$. We run the test in R.

```
>data=c(52,39,49, 33, 58, 61, 44, 221, 133, 289, 211)
>t.test(data, mu=125, alternative="greater")


     One Sample t-test


   data:  data
t = -0.6138, df = 10, p-value = 0.7235
alternative hypothesis:  true mean is greater than 125
```

Thus we get the p value as .7235 and the value of the t statistic as -.6138 with 10 degrees of freedom. The tabulated value of a t statistic with 10 degrees of freedom is 1.81 at 5% level of significance. Thus we accept the null hypothesis at 5% significance level. Therefore, the evidence is not enough to reject the null hypothesis. Consequently, the evidence is not sufficient to conclude that there is an increase in the true mean weekly sales.
But these results are not final because we have not yet checked the assumptions required to perform a t test. Specifically, results from t-tests are valid provided

1. Observations are drawn from a normal parent population, or
2. The sample size is sufficiently large (say, at least 30).

In our example, we have only 11 observations and hence t test is valid if we can show that the underlying distribution is normal.
The simplest descriptive method to check the shape of the data is to plot the histogram. If the resulting histogram is symmetric and bell shaped, the underlying distribution is taken as normal. For the above data we plot the histogram below:

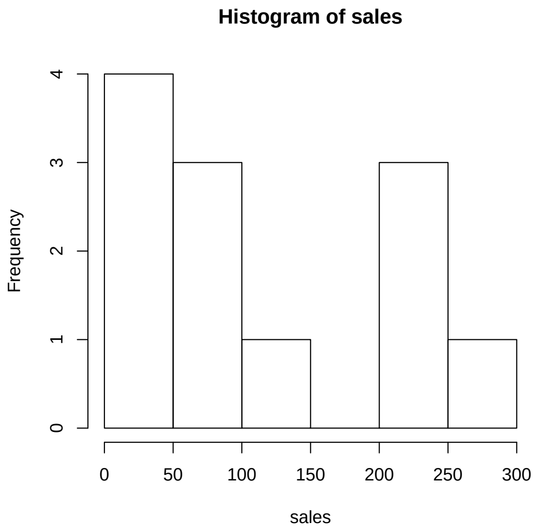


Figure 1: The histogram of data

Clearly, the histogram of sales data is far from being symmetric. This indicates the non normal nature of the underlying distribution of the sales data.
However, for confirmation, we further use a normal Q-Q plot. Normal Q-Q plot is a simple and efficient exploratory technique, where the quantiles of the observed standardised data set are plotted against the corresponding quantiles of a standard normal distribution. If the underlying distribution is standard normal, the result should be a plot of points over a straight line. The normal Q–Q plot for the stanardised version of the sales data is given below.

We find from the plot that the observed quantiles are too far from those of a standard normal distribution and hence the assumption of normality is not reasonable. Thus t test is not trustworthy for the sales data set.
This particular example exhibits the limitations of a parametric procedure in small samples, especially, when the underlying distributional assumption is not satisfied. The above data does not satisfy the assumptions required to perform a t test. As a result, here application of t test is forceful and hence the results from such a test is inconclusive even with a high p value.

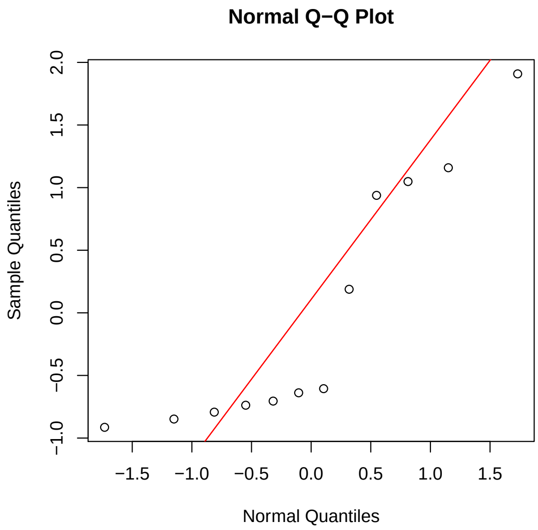


Figure 2: Normal Q–Q plot

Thus there are situations in which parametric assumptions are violated and alternative(i.e. nonparametric) methods of analysis becomes the most appropriate and meaningful.

Nonparametric Inference: Module 2[1]

What we provide in this module

- Different inferential set up
- Functionals
- Plug-in estimators
- Examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A short recap

So far we have learnt that data can be non normal in real practice. Traditional parametric methods then fail to describe the features of the data set. If forcefully the normality assumptions are made, the results can lead to severe error, especially in small sample cases. Then an alternative inferential method is adopted, which does not require any particular distributional assumption unlike parametric methods. These methods are called Non Parametric methods.

2 The set up

Here the data at hand is $\mathbf{x} = (x_1, x_2, ..., x_n)$, where $x_i \in (-\infty, \infty)$ is the observed value of a random variable $X_i, i = 1, 2, .., n$. $F(\mathbf{x}) = P(X_1 \leq x_1, ..., X_n \leq x_n)$ is the distribution function(DF) from which $\mathbf{x}$ is generated. The components of $\mathbf{X}$ are not in general independent, but we assume independence. i.e, $F(\mathbf{x}) = \prod_{i=1}^{n} F_i(x_i)$ for marginal DFs $F_i, i = 1, ..n$. The pair $(\mathbf{X}, F)$ generates the induced probability space $(\mathcal{X}, \mathcal{B}, P)$, where $\mathcal{X}$ is the sample space, $\mathcal{B}$ is the σ field generated by the all possible subsets of $\mathcal{X}$, and$P$ is the induced probability measure. Suppose $p(\mathbf{x})$ is the generalised density of $\mathbf{X}$. In any situation, $p(\mathbf{x})$ is either partially or completely unknown. In statistical inference, we try to give some idea about $p(\mathbf{x})$ or some suitable function or functional of it.

3 Different inferential set up

Depending on the knowledge of $p(\mathbf{x})$(i.e. whether completely or partly unknown), inferential set ups are either-**i. Parametric**,**ii. Nonparametric** and **iii. Semiparametric**

4 Classification

Suppose $p(\mathbf{x})$ is partially known. We can write $p(\mathbf{x}) = \mathbf{p}(\mathbf{x}, \theta), \theta \in \boldsymbol{\Omega}$, where p is functionally known. θ (numerical or abstract valued) is an unknown quantity indexing p, often called a labelling parameter. Ω is the space of all possible values of θ. It is called the parameter space.

Example of parametric set up

1 If θ is real or vector valued, the set up is parametric. For example, suppose X_i are iid observations from a Bernoulli distribution with mean $\pi \in (0, 1)$. Then $\theta = \pi$ is unknown and real valued. $p(\mathbf{x}, \theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \theta^{\mathbf{x_i}}(\mathbf{1} - \theta)^{\mathbf{1-x_i}}$ is a known functional form and $\theta \in (0, 1)$ real valued. Thus the set up is parametric.

2 Suppose X_i are iid observations from a $N(\mu, \sigma^2)$ distribution, $\mu \in R$ and $\sigma > 0$. Then $\theta = (\mu, \sigma)$ is unknown and vector valued. Here $p(\mathbf{x}, \theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{\mathbf{1}}{\sigma\sqrt{\mathbf{2}\pi}}\mathbf{e}^{-\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}}(\mathbf{x_i}-\mu)^{\mathbf{2}}}$ is a known functional form and $\theta \in (-\infty, \infty)X(0, \infty) \subset R^2$. Thus the set up is parametric.

Example of semi parametric set up

1 If θ is partly real or vector valued and partly abstract valued, the set up is semi-parametric. For example, suppose X_i are iid observations from a density $f(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}exp(-\frac{(x-\mu)^2}{2}) + \frac{1}{2}g(x)$ where g is unknown but known to be continuous. Then $p(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i)$ is a partly known functional form and $\theta = (\mu, g)$ is partly abstract valued and partly real valued. Thus the set up is semiparametric.

2 Consider the regression set up with one non stochastic covariate. Suppose X_i are independent observations from the continuous density $f_i, i = 1, 2, .., n$ with $E(X_i) = \alpha + \beta z_i$ and $Var(X_i) = \sigma^2$. z_i are known quantities, $i = 1, 2, .., n$. Then $\theta = (\alpha, \beta, \sigma, f_1, .., f_n)$ is partly abstract valued and partly real valued and hence the set up is semiparametric.

Example of non parametric set up

1 If θ is completely abstract valued, the set up is nonparametric.

Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown DF $F(x)$, where The functional form of F is unknown but known to be continuous. Then $\theta = F$ is unknown and abstract valued. $\Omega =$ is the class of all absolutely continuous DFs. Thus the set up is nonparametric.

2 Consider two independent samples $\{X_i, i = 1, .., m\}$ and $\{Y_j, j = 1, .., n\}$, where X_i are iid observations from an unknown density $f(x)$ Y_j are iid observations from an unknown density $g(x)$, where f and g are both unknown but known to be continuous. Then $\theta = (f, g)$ is unknown and abstract valued. Thus the set up is nonparametric.

5 Concept of functional

In a nonparametric set up, the basic assumption is the continuity of the underlying distribution F. The unknown quantity of interest (parametric function in parametric set up) is defined in terms of a functional. If $\mathcal{F}$ is the class of all absolutely continuous distributions, then a functional $\theta_{\mathrm{F}} = \theta(F)$ is a real valued function defined for every F $\in \mathcal{F}$ Then $\theta_{\mathrm{F}} : \mathcal{F} \to R$ is a mapping from the abstract space to real line. In nonparametric inference, the objective is to learn the value of $\theta_{\mathrm{F}} = \theta(F)$ for a known functional θ based on iid observations from an unknown DF F.

Example of functionals

Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown DF $F(x)$, where F is unknown but known to be continuous. Few possible functionals are:

1. $\theta_{\mathrm{F}} = E_F I(X \leq a)$, for some known a, where $\mathcal{F}$ is the class of all absolutely continuous DFs.

2. $\theta_{\mathrm{F}} = E_F(X)$, where $\mathcal{F} = \{\mathrm{F} : E_F(|X|) < \infty\}$

3. $\theta_{\mathrm{F}} = E_F(X^2)$, where $\mathcal{F} = \{\mathrm{F} : E_F(X^2) < \infty\}$

Linear functional

Each of the functionals can be expressed as $E_F\psi(X)$, for some ψ.

For examples 2 and 3, $\psi(X) = X$ and X^2.

For example 1, $\psi(X) = I(X \leq a)$, where $I(\ )$ is the indicator function. Then $\theta(\alpha F_1 + (1-\alpha)F_2) = \alpha\theta_{\mathrm{F}_1} + (1-\alpha)\theta_{\mathrm{F}_2}$ for any $0 < \alpha < 1$. The above functionals are called linear functionals.

Nonlinear functionals

Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown DF $F(x)$, where F is unknown but known to be continuous. Then the following are also valid functionals:

1. $\theta_{\mathrm{F}} = \{E_F(X)\}^2$, where $\mathcal{F} = \{\mathrm{F} : E_F(|X|) < \infty\}$

2. $\theta_{\mathrm{F}} = \mathrm{Var}_{\mathrm{F}}(X)$, where, $\mathcal{F} = \{\mathrm{F} : E_F(X^2) < \infty\}$

The above functionals are nonlinear functionals as $\theta(\alpha F_1(1-\alpha)F_2) \neq \alpha\theta_{\mathrm{F}_1} + (1-\alpha)\theta_{\mathrm{F}_2}$ for $0 < \alpha < 1$.

Further examples

Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown DF $F(x)$, where F is unknown but known to be continuous. Then the following are also nonlinear functionals:

1. $\theta_{\mathrm{F}} = F^{-1}(\frac{1}{2})$, the median , where, $\mathcal{F}$ is the class of all absolutely continuous DFs.

2. $\theta_{\mathrm{F}} = \frac{F^{-1}(\frac{3}{4}) - F^{-1}(\frac{1}{4})}{2}$, the quartile deviation, where, $\mathcal{F}$ is the class of all absolutely continuous DFs.

A critical example

Suppose $\mathcal{F}$ is class of all absolutely continuous DF with finite expectation. Then the question is what will be the type of the kernel $\theta_{\mathrm{F}} = \{E_F(X)\}^2$?

Take $F_i \in \mathcal{F}, i = 1, 2$ Then for any $\alpha \in (0, 1)$,

$E_{\alpha F_1 + (1-\alpha) F_2}(X) = \alpha E_{F_1}(X) + (1 - \alpha) E_{F_2}(X)$.

But $\{E_{\alpha F_1 + (1-\alpha) F_2}(X)\}^2 \neq \alpha \{E_{F_1}(X)\}^2 + (1 - \alpha)\{E_{F_2}(X)\}^2$.

Therefore, θ_{F} is not a linear functional.

Plug in estimator: Analogue to statistic

Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown DF $F(x)$, where F is unknown but known to be continuous. F_n is the empirical DF based on the data. $\theta(F_n)$ is the plug-in estimator corresponding to the functional θ(F)

1. $E_{F_n}\psi(X)$ is the plug-in estimator corresponding to the functional $E_F\psi(X)$.
2. $F_n^{-1}(\frac{1}{2})$, the sample median is the plug-in estimator corresponding to $F^{-1}(\frac{1}{2})$.

A misleading example

Suppose $\mathcal{F}$ is class of all absolutely continuous DF with symmetry at the origin. Then what will be the type of the functional $\theta_F = P(X > 0)$?

At a first look it appears as a linear functional. But symmetry at the origin gives $\theta_F = \frac{1}{2}$. Therefore, θ_F is not a functional.

Nonparametric Inference: Module 3[1]

What we provide in this module

- Estimability
- Kernels and Symmetry
- Sum and product of kernels
- U statistic
- Few examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A short recap

So far we have learnt the following

- Functionals are the analogues of parameter of interest.
- Statistical functionals are the analogues of statistic.
- Functionals are linear and non linear.
- Functionals and statistical functionals form the basis for nonparametric inference.

2 Inferential problems in nonparametrics

As in the parametric inference, we have three different types of inferential problems:

i.Estimation

ii.Hypothesis Testing

iii. Confidence Interval estimation

In this module we start with nonparametric estimation.

3 Estimability

We have already introduced functional and a statistical functional. Suppose $X_1, X_2, ..., X_n$ are iid observations from F and θ_{F} is a functional defined on $\mathcal{F}$, the class of all absolutely continuous DFs.

Then θ_{F} is said to be estimable if there exists a $\phi(X_1, X_2, ..., X_k)$,k$\leq$ n, such that,

$E_F\{\phi(X_1, X_2, ..., X_k)\} = \theta_F$, for all F $\in \mathcal{F}$.

If the above holds for some ϕ and k, θ_F is called a regular functional.

ϕ is called a kernel for estimation of $\theta_{\mathcal{F}}$. Minimum value of k, ensuring the above is called the degree of the kernel.

4 Sum / Difference of the kernels

If ϕ_i is a kernel of degree k_i for $\theta_i(F), i = 1, 2$, then we are interested to know degree of $\phi_1 \pm \phi_2$. Observe that $\phi_1 \pm \phi_2$ is a kernel for $\theta_1(F) \pm \theta_2(F)$ Few observations can be common between ϕ_1 and ϕ_2.

Assume that $k_1 < k_2$. Then $\phi_1 \pm \phi_2$ is based on at most k_2 observations. If $k_1 > k_2$, then $\phi_1 \pm \phi_2$ is based on at most k_1 observations. In any case $\phi_1 \pm \phi_2$ involves at most $max(k_1, k_2)$ observations. So, degree of $\phi_1 + \phi_2$ is at most $max(k_1, k_2)$.

Generalization: Degree of $\sum_{i=1}^{s} \pm\phi_i$ can not exceed $max(k_1, k_2, .., k_s)$

5 Product of kernels

If ϕ_i is a kernel of degree k_i for $\theta_i(F), i = 1, 2$, then we are interested to know the degree of $\phi_1\phi_2$. Suppose there are some common observations in ϕ_1 and ϕ_2. Then ϕ_1 and ϕ_2 are, in general, dependent. Thus product of them is not a kernel for $\theta_1(F)\theta_2(F)$. Suppose $\theta_1(F)\theta_2(F)$ is the functional of interest. Then $\phi_1\phi_2$ is a kernel for $\theta_1(F)\theta_2(F)$, if they are based on different sets of observations. Assume that the sets of observations corresponding to k_1 and k_2 are different Then $\phi_1\phi_2$ involves $k_1 + k_2$ distinct observations. Degree of $\phi_1\phi_2$ is $k_1 + k_2$.

Generalization: Degree of the kernel $\prod_{i=1}^{s} \pm\phi_i$ is $\sum_{i=1}^{s} k_i$ for the functional $\prod_{i=1}^{s} \theta_i(F)$.

6 Symmetric kernels

A kernel $\phi(x_1, x_2, ..., x_k)$ is called symmetric if, $\phi(x_1, x_2, ..., x_k) = \phi(x_{i_1}, x_{i_2}, ..., x_{i_k})$ for any permutation $(i_1, i_2, ..., i_k)$ of $\{1, 2, ..., k\}$.

1 $\sum_{i=1}^{n} x_i$ or any function of $\sum_{i=1}^{n} x_i$ are symmetric kernels.

2 Median, quantiles or any function of them does not depend on the order of the data and hence are symmetric kernels.

3 $I(x_1 + x_2 > 1)$ is a symmetric kernel.

7 Asymmetric kernels

A kernel $\phi(x_1, x_2, ..., x_k)$ is asymmetric if it is not permutation invariant, i.e. $\phi(x_{i_1}, x_{i_2}, ..., x_{i_k})$ and $\phi(x_1, x_2, ..., x_k)$ are not the same for some permutation $(i_1, i_2, ..., i_k)$ of $\{1, 2, ..., k\}$.

1 $\phi(x_1, x_2) = x_1^2 - x_1 x_2$ is an asymmetric kernel as $\phi(x_1, x_2) \neq \phi(x_2, x_1)$.

2 $x_1 - x_2$ and $\frac{x_1}{x_2}$ are further examples of asymmetric kernels.

3 As another example, $\phi(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - x_1 x_3$ is asymmetric as $\phi(x_1, x_2, x_3) = \phi(x_3, x_2, x_1)$ but $\phi(x_1, x_2, x_3) \neq \phi(x_3, x_1, x_2)$.

A useful result

Result: For every regular functional, there exists a symmetric kernel.

Proof: Suppose $X_1, .., X_n$ are iid observations from F and θ_F is a regular functional. If $\phi(x_1, x_2, ..., x_k)$ is an asymmetric kernel then $E\phi(x_1, x_2, ..., x_k) = \theta_F$. As the observations are iid,

$$E\phi(X_{i_1}, X_{i_2}, ..., X_{i_k}) = \theta_F$$

for any permutation $(i_1, i_2, ..., i_k)$ of $\{1, 2, ..., k\}$. We construct the symmetric function $\phi_s(x_1, x_2, ..., x_k) = \frac{1}{k!} \sum_{i_1, i_2, ..., i_k} \phi(x_{i_1}, x_{i_2}, ..., x_{i_k})$. Then
$E\phi_s(X_1, X_2, ..., X_k) = \theta_F$.

Without any loss of generality, kernels can be taken as symmetric.

8 U statistic

Let $X_1, X_2, ..., X_n$ be iid observations from F. For any kernel ϕ of degree k, U statistic is defined as

$$U_n = \frac{(n-k)!}{k!} \sum_{1 \leq i_1 \neq i_2 \neq .. \neq i_k \leq n} \phi(X_{i_1}, X_{i_2}, ..., X_{i_k}).$$

If ϕ is symmetric, U_n takes the form

$$U_n = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < .. < i_k \leq n} \phi(x_{i_1}, x_{i_2}, ..., x_{i_k}).$$

We note that

- U_n is permutation invariant
- U_n is based on all observations.
- For further discussion, we consider ϕ as symmetric.

Example 1

We consider the set up $\mathcal{F} = \{F : E_F|X| < \infty\}$, $\theta_F = E(X_1)$

Take $\phi(X_1) = X_1$, then $E\phi = \theta_F$ and ϕ is a symmetric kernel of degree $k = 1$.

Corresponding U statistic is:

$$U_n = \tfrac{1}{n} \sum_{1 \leq i_1 \leq n}^{n} \phi(X_{i_1}) = \bar{X}_n$$

Thus we get the usual unbiased estimator.

Example 2

We consider the set up $\mathcal{F}$=class of all absolutely continuous DFs and $\theta_F = P(X_1 \leq a)$, a is known. We observe that $\theta_F = EI(X_1 \leq a)$. This suggests to take $\phi(X_1) = I(X_1 \leq a)$, as a kernel ϕ is a symmetric kernel of degree $k = 1$. Then the corresponding U statistic is:

$$U_n = \tfrac{1}{n} \textstyle\sum_{i_1=1}^{n} I(X_{i_1} \leq a)$$

We get the usual unbiased estimator-proportion of observations less than a.

Example 3

We consider $\mathcal{F} = \{F : E_F|X| < \infty\}$ and $\theta_F = \{E(X_1)\}^2$. Since, $E(X_1X_2) = E(X_1)E(X_2) = \{E(X_1)\}^2$, we take $\phi(X_1, X_2) = X_1X_2$, then $E\phi = \theta_F$, where $\phi(X_{i_1}, X_{i_2}) = \{\bar{X_n}\}^2 - \frac{s^2}{n}$ with s^2 as the sample variance with divisor $n-1$.

Example 4

We consider the set up $\mathcal{F} = \{F : E_FX^2 < \infty\}$, $\theta_F = Var(X_1)$ Observe that $Var(X_1) = E(X_1^2) - \{E(X_1)\}^2$. Thus θ_F is a difference of two quantities. For the first quantity, a kernel is X_1^2 and a kernel for the second quantity is X_1X_2.

We take the difference of these two as a kernel. That is we take $\phi_0(X_1, X_2) = X_1^2 - X_1X_2$, as a kernel. But ϕ_0 is asymmetric kernel of degree $k = 2$. Corresponding symmetric kernel is,

$\phi(X_1, X_2) = \frac{1}{2}\{\phi_0(X_1, X_2) + \phi_0(X_2, X_1)\} = \frac{1}{2}(X_1 - X_2)^2$

Corresponding U statistic is

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 \le i_2 \le n}^{n} \phi(X_{i_1}, X_{i_2}).$$

We can simplify U as

$$\begin{aligned}
U_n &= \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 < i_2 \le n} \frac{(X_{i_1} - X_{i_2})^2}{2} \\
&= \frac{1}{\binom{n}{2}} \frac{1}{2} \left[\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \frac{(X_{i_1} - X_{i_2})^2}{2} \right] \\
&= \frac{1}{2n(n-1)} \left[n \sum_{i_1=1}^{n} X_{i_1}^2 + n \sum_{i_2=1}^{n} X_{i_2}^2 - 2(\sum_{i_1=1}^{n} X_{i_1})(\sum_{i_2=1}^{n} X_{i_2}) \right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right] = s^2
\end{aligned}$$

Example 5

We consider the set up, $\mathcal{F} = \{F : E_FX^2 < \infty\}$, $\theta_F = E(X_1^2)$

Observe that $\theta_F = Var(X_1) + \{E(X_1)\}^2$ For the first component, a kernel is $X_1^2 - X_1X_2$

with degree 2 and for the second component, a kernel is X_1X_2 with degree 2. For θ_F, the kernel is the sum X_1^2, which has degree 1.
Corresponding U statistic is

$$U_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

We at once observe the following:
The kernel is a sum of two kernels, each of degree 2. But the resultant kernel is of degree 1. Thus the degree of the sum of two kernels can be strictly less than 2. The inequality can be strict in Result. Again, one of the kernels is symmetric and the other asymmetric. But the resultant kernel is symmetric. In particular the sum of two symmetric kernels are always symmetric.

Example 6

We consider the set up $\mathcal{F}$=class of all bivariate absolutely continuous DFs. $(X_i, Y_i), i = 1, 2, .., n$ are iid observations from $F \in \mathcal{F}$ and $\theta_F = E_F(X_1Y_1), Var(X+Y)$.
Define a new variable $Z_i = X_iY_i, i = 1, 2, .., n$. Then Z_i are iid observations from some univariate absolutely continuous DF G. Thus θ_F reduces to $\theta_G = E_G(Z_1)$. Then the corresponding U statistic is $U_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$

Example 7

We consider the set up $\mathcal{F}$=class of all bivariate absolutely continuous DFs $(X_i, Y_i), i = 1, 2, .., n$ are iid observations from $F \in \mathcal{F}$ and $\theta_F = Var(X+Y)$.
Define a new variable $Z_i = X_i + Y_i, i = 1, 2, .., n$. Then Z_i are iid observations from some univariate absolutely continuous DF G. Thus θ_F reduces to $\theta_G = Var_G(Z_1)$. Hence the corresponding U statistic is $U_n = \frac{1}{n-1}\left[\sum\limits_{i=1}^{n} Z_i^2 - n\bar{Z}^2\right]$

Nonparametric Inference: Module 4[1]

What we provide in this module

- Sufficiency of order statistics
- Improvement using U statistic
- U is MVUE
- Exact variance of U
- Bounds for variance
- Few examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A few concepts

Suppose $\mathbf{X} = (X_1, X_2, .., X_n)$ are iid observations from an absolutely continuous distribution F. $\mathcal{F}$ is the class of all absolutely continuous DFs. $\mathbf{Y} = (X_{(1)}, X_{(2)}, .., X_{(n)})$ is the full set of order statistic corresponding to $\mathbf{X}$. $\mathcal{P}_n$ is the set of $n!$ permutations of the first n natural numbers. The joint pdf of $(X_1, X_2, .., X_n)$ given the full set of order statistics is

$$f(\mathbf{x}|\mathbf{y}) = \frac{1}{n!}, \mathbf{x} \in \mathcal{P}_n.$$

Thus $\mathbf{Y}$ is sufficient for $\mathcal{F}$.

2 U statistics as conditional expectation

Suppose θ_{F} is a functional defined on $\mathcal{F}$. Also suppose $\phi(X_1, X_2, ..., X_k)$,k$\leq$ n is a kernel for θ_F. That is, $\phi(X_1, X_2, ..., X_k)$, is an unbiased estimator of θ_F. Then

$$E\{\phi(X_1, X_2, ..., X_k)|\mathbf{Y}\} = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < .. < i_k \leq n} \phi(X_{i_1}, X_{i_2}, ..., X_{i_k})$$

The RHS is nothing but the conventional U statistic with kernel ϕ of degree k.

3 An improved estimator from an Unbiased estimator

Let $T_n(X_1, X_2, ..., X_n)$ be an UE of θ_F. Corresponding U statistic is,
$U_n = \frac{1}{n!} \sum_{(i_1, i_2, ..., i_n) \in \mathcal{P}_n} T_n(X_{i_1}, X_{i_2}, ..., X_{i_n}) = E(T_n(X_1, X_2, ..., X_n)|\mathbf{Y})$.
Then $E(U_n^2) = E\{E(T_n|\mathbf{Y})\}^2 \leq E\{E(T_n^2|\mathbf{Y})\} = ET_n^2$. Since $E(U_n) = E(T_n) = \theta_F$,

$$Var(U_n) \leq Var(T_n)$$

Thus U_n is an improvement over the UE T_n.

4 U statistic & MVUE

Suppose $\mathcal{F}$ is the class of all absolutely continuous DFs. Then $\mathbf{Y} = (X_{(1)}, X_{(2)}, .., X_{(n)})$ is complete sufficient for $\mathcal{F}$(see, Fraser,1957). U_n is symmetric in the observations and hence is a function of the full set of order statistics. Thus U_n is Unbiased and function of complete sufficient statistics. Therefore U_n is MVUE for θ_F

5 Exact variance of U statistic

Suppose $\phi(X_1, X_2, ..., X_k)$,k$\leq$ n is a symmetric kernel for θ_F and $\mathcal{F}$ is the class of distributions with finite $V_F(\phi(X_1, X_2, ..., X_k)$. Define

$\phi_m(x_1, x_2, .., x_m) = E\phi(X_1, X_2, .., X_m, X_{m+1}, .., X_k | X_1 = x_1, .., X_m = x_m)$

$= E\phi(x_1, x_2, .., x_m, X_{m+1}, .., X_k),\ 1 \leq m \leq k$ and

$\sigma_m^2 = Var[\phi_m(X_1, X_2, .., X_m)]\ , 1 \leq m \leq k$

Then $Var(U_n)$ can be expressed as $\sum_{m=1}^{k} p_{n,k}(m)\sigma_m^2\ p_{n,k}(m) = \frac{\binom{k}{m}\binom{n-k}{k-m}}{\binom{n}{k}},\ 1 \leq m \leq k$

A related observation

With the already introduced notations, the smallest and the largest members of the set $\{\sigma_1^2, \sigma_2^2, .., \sigma_k^2\}$ are, respectively σ_1^2 and σ_k^2.

Applying Jensen inequality, we get

$$\begin{aligned} E\phi_{m+1}^2(X_1, X_2, .., X_{m+1}) &= EE\{\phi_{m+1}^2(X_1, X_2, .., X_{m+1}) | X_{m+1}\} \\ &\geq E\{E[\phi_{m+1}(X_1, X_2, .., X_{m+1}) | X_{m+1}]\}^2 \end{aligned}$$

Since $E[\phi_{m+1}(X_1, X_2, .., X_{m+1}) | X_{m+1}] = \phi_m(X_1, X_2, .., X_m)$, we have $E\phi_{m+1}^2(X_1, X_2, .., X_{m+1}) \geq E\phi_m^2(X_1, X_2, .., X_m)$.

As $\sigma_m^2 = E\phi_m^2(X_1, X_2, .., X_m) - \theta_F^2$, we get $\sigma_{m+1}^2 \geq \sigma_m^2$ for any m.

6 Derivation of $Var(U_n)$

First of all note that U_n is a sum of dependent variables. Therefore the covariance terms are not always zero. Thus

$$Var(U_n) = \binom{n}{k}^{-2} \sum_{1\leq i_1<..<i_k\leq n} \sum_{1\leq j_1<..<j_k\leq n} Cov\{\phi(X_{i_1},..,X_{i_k},\phi(X_{j_1},..,X_{j_k})\}$$

We start with proving a related lemma.

Lemma: If m is the number of members common between the sets $\{i_1, i_2, .., i_k\}$ and $\{j_1, j_2, .., j_k\}$, we have $Cov\{\phi(X_{i_1},..,X_{i_k}),\phi(X_{j_1},..,X_{j_k})\} = \sigma_m^2$, where $\sigma_0^2 = 0$.

Proof: Since $X_i, i = 1, 2, .., n$ are iid , we have $Cov\{\phi(X_{i_1},..,X_{i_k}),\phi(X_{j_1},..,X_{j_k})\} = 0$ or $\neq 0$ as $m = 0$ or $m > 0$. Consequently, for $m \neq 0$,

$E\{\phi(X_{i_1},..,X_{i_k}\phi(X_{j_1},..,X_{j_k})\}$

$= E\phi(X_1,..,X_m,X_{m+1},..,X_k)\phi(X_1,..,X_m,X_{k+1},..,X_{2k-m})$

Now, given $X_1,..,X_m$, the above two kernels are independent. Then by definition

$E\{\phi(X_1,..,X_m,X_{m+1},..,X_k)|X_1,..,X_m\} = \phi_m(X_1,..,X_m)$

$E\{\phi(X_1,..,X_m,X_{k+1},..,X_{2k-m})|X_1,..,X_m\} = \phi_m(X_1,..,X_m)$ Thus

$E\{\phi(X_{i_1},..,X_{i_k}\phi(X_{j_1},..,X_{j_k})\} = E\phi_m^2(X_1,..,X_m)$

Hence,

$Cov\{\phi(X_{i_1},..,X_{i_k},\phi(X_{j_1},..,X_{j_k})\}$

$=E\phi(X_{i_1},..,X_{i_k})\phi(X_{j_1},..,X_{j_k}) - \theta_F^2$

$=E\phi_m(X_1,..,X_m)^2 - \theta_F^2$

$=\sigma_m^2$

And the lemma follows. Next we continue to the result for variance.

In view of the lemma, the contribution of each covariance term is either 0 or σ_m^2. For a specific $m, 1 \leq m \leq k$, the coefficient of σ_m^2 is the number of choices of the sets $\{i_1, i_2, .., i_k\}$ and $\{j_1, j_2, .., j_k\}$ having m elements in common. The number of such choices is $\binom{n}{k}\binom{k}{m}\binom{n-k}{k-m}$.

Because, there are $\binom{n}{k}$ ways of choosing a particular $\{i_1, i_2, .., i_k\}$ and then $\binom{k}{m}$ ways of choosing c elements from them. Again, there are $\binom{n-k}{k-m}$ ways of selecting the remaining $k-m$ elements of $\{j_1, j_2, .., j_k\}$ from the remaining $n-k$ numbers. Finally,for a particular m, there are $\binom{n}{k}\binom{k}{m}\binom{n-k}{k-m}$ terms with covariance σ_m^2. Then $Var(U_n) = \binom{n}{k}^{-2} \sum_{m=1}^{k} \binom{n}{k}\binom{k}{m}\binom{n-k}{k-m}\sigma_m^2$. A simplification gives the desired expression of variance.

7 An alternative look at the variance

Note that $p_{n,k}(m),\ 1 \leq m \leq k$ is the pmf of a Hypergeometric distribution. Define M, a random variable with $P(M = m) = p_{n,k}(m)$ Then, we have the following representation of U statistic U_n:

$E(U_n|M) = \theta_F$ and $Var(U_n|M) = \sigma_M^2$.

It follows from such a representation, $E(U_n) = EE(U_n|M) = \theta_F$ and

$$\begin{aligned} Var(U_n) &= E\{Var(U_n|M)\} + Var\{E(U_n|M)\} \\ &= E\sigma_M^2 \\ &= \sum_{m=1}^{k} \sigma_m^2 P(M = m) \\ &= \sum_{m=1}^{k} \sigma_m^2 p_{n,k}(m) \end{aligned}$$

8 Exact bounds for $Var(U_n)$

We have already obtained that

$\sigma_1^2 \leq \sigma_2^2 \leq ... \leq \sigma_k^2$ and $Var(U_n) = \sum_{m=1}^{k} \sigma_m^2 p_{n,k}(m)$

For fixed $n, k,\ p_{n,k}(m),\ 1 \leq m \leq k$ is the pmf of a Hypergeometric distribution. Then it follows easily that $\sigma_1^2 \leq Var(U_n) \leq \sigma_k^2$.

9 Finding $Var(U_n)$

Assume that $X_i, i = 1, 2, .., n$ are iid observations from F.

Example 1: $\theta_F = E_F(X_1)$, $\phi(X_1) = X_1$, $k = 1$, $\sigma_1^2 = Var\phi(X_1) = Var(X_1)$, $Var(U_n) = Var(X_1)/n$

Example 2: $\theta_F = E_F(X_1^r)$, $\phi(X_1) = X_1^r$, $k = 1$ r is a known positive integer. Assume that $E(X_1^s) = \mu_s^{'} < \infty$ for $s = 2r$. $\sigma_1^2 = Var\phi(X_1) = Var(X_1^r) = \mu_{2r}^{'} - \mu_r^{'2}$. Thus $Var(U_n) = (\mu_{2r} - \mu_r^{'2})/n$

Example 3: $\theta_F = P_F(X_1 \leq x_0)$, $\phi(X_1) = I(X_1 \leq x_0)$, $k = 1$

Assume that x_0 is known, then $\sigma_1^2 = Var\phi(X_1) = VarI(X_1 \leq x_0)$. Thus $Var(U_n) = \theta_F(1 - \theta_F)/n$.

Example 4: $\theta_F = \{E_F(X_1)\}^2$, $\phi(X_1, X_2) = X_1X_2$, $k = 2$. Assume $E(X_1) = \mu$ and $Var(X_1) = \sigma^2 < \infty$

$\phi_1(X_1) = E\{\phi(X_1, X_2)|X_1\} = \mu X_1$

$\sigma_1^2 = Var\phi_1(X_1) = \mu^2\sigma^2$

$\phi_2(X_1, X_2) = \phi(X_1, X_2) = X_1X_2$

$\sigma_2^2 = Var\phi_2(X_1, X_2) = Var(X_1X_2) = \sigma^4 + 2\mu^2\sigma^2$

$Var(U_n) = \frac{\binom{2}{1}\binom{n-2}{1}}{\binom{n}{2}}\mu^2\sigma^2 + \frac{\binom{2}{2}\binom{n-2}{0}}{\binom{n}{2}}(\sigma^4 + 2\mu^2\sigma^2)$

Example 5: $\theta_F = Var(X_1)$, $\phi(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$, $k = 2$

Assume $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$, $E(X_1 - \mu)^4 = \mu_4 < \infty$.

$\phi_1(X_1) = E\{\phi(X_1, X_2)|X_1\} = \sigma^2/2 + (X_1 - \mu)^2/2$

$\sigma_1^2 = Var\phi_1(X_1) = Var\{(X_1 - \mu)^2/2\} = (\mu_4 - \sigma^4)/4$

$\phi_2(X_1, X_2) = \phi(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$

$\sigma_2^2 = Var\phi_2(X_1, X_2) = (\mu_4 + \sigma^4)/2$

$Var(U_n) = \frac{\binom{2}{1}\binom{n-2}{1}}{\binom{n}{2}}(\mu_4 - \sigma^4)/4 + \frac{\binom{2}{2}\binom{n-2}{0}}{\binom{n}{2}}(\mu_4 + \sigma^4)/2 = (\mu_4 - (n-3)\sigma^4/n - 1)/n$

Nonparametric Inference: Module 5[1]

What we provide in this module

- Consistency of U statistic
- Asymptotic normality of U statistic
- Degeneracy of U statistic
- Few examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Large Sample Properties

Unbiasedness and possessing minimum variance are the small sample properties of an estimator. To judge the performance of an estimator for large sample sizes, one uses Consistency and asymptotic normality. Consistency gives the limiting value whereas asymptotic normality specifies the rate of convergence to the limiting value. A consistent estimator having asymptotic normality is known as Consistent Asymptotic Normal(CAN). U statistic is a CAN estimator.

Result: Suppose U_n is an U statistic with kernels of degree k and $\sigma_1^2 > 0$ and $\sigma_i^2 < \infty,\ i \geq 2$. Then $lim_{n\to\infty} Var(\sqrt{n}U_n) = k^2\sigma_1^2$.

Proof: For any a, $n\frac{\binom{n-k}{k-a}}{\binom{n}{k}} = \frac{k!}{(k-a)!}n\frac{\prod_{j=k}^{k-a-1}(n-j)}{\prod_{j=0}^{k-1}(n-j)}$. Then the denominator is the product of k-a+1 terms while the denominator involves k terms only. Each term is linear in n. Thus for $a > 1$, the RHS goes to zero whereas for a=1, it goes to k. Thus $Var(\sqrt{n}U_n) = k^2\sigma_1^2 + o(n^{-1})$.Hence, as $n \to \infty$, $Var(\sqrt{n}U_n) \to k^2\sigma_1^2$.

2 Consistency of U statistic

Now we come to the proof of consistency of U statistic. First of all, by definition, $E(U_n) = \theta_F$. Consider the representation $Var(U_n) = Var(\sqrt{n}U_n)\frac{1}{n}$. Then, as $n \to \infty$, the first term approaches $k^2\sigma_1^2$ by the previous result. The second term converges to zero. Thus $Var(U_n) \to 0$ as $n \to \infty$. Hence consistency follows.

3 Asymptotic Distribution of U statistic

Result: Provided $\sigma_k^2 > 0$, as $n \to \infty$,

$$\sqrt{n}(U_n - \theta_F) \to N(0, k^2\sigma_1^2)$$

in distribution.

Proof: We start with the fact that for $\sigma_1^2 = 0$, $U_n = \theta_F$ with probability and hence assume $\sigma_1^2 > 0$ for further development.
First of all, we note that U_n is not a sum of iid random random variables so that the large sample distribution can not be obtained by a straightforward application of Central Limit Theorem.
Consider the representation,

$$\sqrt{n}(U_n - \theta_F) \quad = \tfrac{k}{\sqrt{n}} \textstyle\sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\} + \epsilon_n \cdots (*),$$

where

$$\epsilon_n = \sqrt{n}(U_n - \theta_F) - \frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\}.$$

Now $\phi_1(X_i)$ are iid random variables, and hence by Lindeberg-Levy CLT

$$\frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\} \xrightarrow{D} N(0, k^2\sigma_1^2)$$

as $n \to \infty$. Then the proof will follow, if we can establish that $\epsilon_n \xrightarrow{P} 0$ as $n \to \infty$.

Since $E(\epsilon_n) = 0$, we shall show only that $E(\epsilon_n^2) \to 0$. Now

$$\begin{aligned} E(\epsilon_n^2) &= Var\{\sqrt{n}(U_n - \theta_F)\} + Var[\frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\}] \\ &- 2Cov(\sqrt{n}(U_n - \theta_F), \frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\}) \cdots (*) \end{aligned}$$

The first term in the RHS of (*) converges to $k^2\sigma_1^2$. The next term in the RHS of (*) converges again to $k^2\sigma_1^2$. Thus the proof will be complete provided the last term in the RHS also converges to $k^2\sigma_1^2$.

Now

$$Cov(\sqrt{n}(U_n - \theta_F), \frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\})$$
$$= \frac{k}{\binom{n}{k}} \sum_{i=1}^{n} \sum Cov\{\phi(X_{i_1}, .., X_{i_k}), \phi_1(X_i)\}$$

Since the observations are iid, the inside covariance term is nothing but $Cov(\phi(X_1, .., X_k), \phi_1(X_1))$ A simple use of conditioning expresses the covariance as σ_1^2.

Now there are $\binom{n}{1}\binom{n-1}{k-1}$ covariance terms with value σ_1^2. Then we get the desired expression (i.e. $k^2\sigma_1^2$) for the last term in the RHS of (*). As a result $\sqrt{n}(U_n - \theta_F)$ and $\frac{k}{\sqrt{n}} \sum_{i=1}^{n} \{\phi_1(X_i) - \theta_F\}$ have the same asymptotic distribution. Since the latter has the desired asymptotic distribution, the proof concludes.

4 Few Examples-$k = 1$

Assume that $X_i, i = 1, 2, .., n$ are iid observations from F. **Example 1:** $\theta_F = E_F(X_1)$, $\phi(X_1) = X_1$, $U_n = \bar{X}_n$

$\sigma_1^2 = Var\phi(X_1) = Var(X_1)$

$\sqrt{n}(U_n - \theta_F) \xrightarrow{D} N(0, Var(X_1))$

Example 2: $\theta_F = E_F(X_1^r)$, $\phi(X_1) = X_1^r$, $U_n = \frac{1}{n} \sum_{i=1}^{n} X_i^r$

$\sigma_1^2 = Var\phi(X_1) = Var(X_1^r) = \mu_{2r}^{'} - \mu_r^{'2}$

$\sqrt{n}(U_n - \theta_F) \xrightarrow{D} N(0, \mu_{2r}^{'} - \mu_r^{'2})$

5 Few Examples-$k = 2$

Example 3: $\theta_F = \{E_F(X_1)\}^2$, $\phi(X_1, X_2) = X_1X_2$, $U_n = \bar{X}^2 - s^2/n$

Assume $E(X_1) = \mu$ and $Var(X_1) = \sigma^2 < \infty$

$\sigma_1^2 = Var\phi_1(X_1) = \mu^2\sigma^2$

$\sqrt{n}(U_n - \theta_F) \xrightarrow{D} N(0, 4\mu^2\sigma^2)$

Example 4: $\theta_F = Var(X_1)$, $\phi(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$, $U_n = s^2$

Assume $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$, $E(X_1 - \mu)^4 = \mu_4 < \infty$.

$\sigma_1^2 = Var\phi_1(X_1) = Var\{(X_1 - \mu)^2/2\} = (\mu_4 - \sigma^4)/4$

$\sqrt{n}(U_n - \theta_F) \xrightarrow{D} N(0, \mu_4 - \sigma^4)$

6 Degeneracy of U statistic

It may happen that the asymptotic variance of a U statistic is zero. For example, suppose $\theta_F = \{E_F(X_1)\}^2$. Assume $E(X_1) = \mu = 0$ and $Var(X_1) = \sigma^2 < \infty$. Then $\sigma_1^2 = Var\phi_1(X_1) = \mu^2\sigma^2 = 0$. Naturally, $\sqrt{n}(U_n - \theta_F)$ converges to a degenerate distribution. Thus the asymptotic normality does not hold.

Definition: Consider a U statistic with symmetric kernels of degree k. U is said to have a degeneracy of order m if

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 = 0,$$

but $\sigma_{m+1}^2 > 0$. Alternatively, since for a U statistic with kernels of degree k,

$$\sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_k^2,$$

the definition can be restated as

U is said to have a degeneracy of order m if $\sigma_m^2 = 0$ but $\sigma_{m+1}^2 > 0$.

An Example

Consider the example already stated. Naturally, for $\mu = 0$, $\sigma_1^2 = 0$ but $\sigma_2^2 = \sigma^4 > 0$ Then U has degeneracy of order 1. Under the presence of degeneracy, we need to use different normalizing constant to reach a non-degenerate limiting distribution. To obtain the limiting distribution under degeneracy, we start from

$$nU_n = \frac{n}{n-1}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\}^2 - \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} X_i^2 \cdots (*)$$

The first term in the RHS of (*) converges in distribution to a $\sigma^2\chi_1^2$ distribution. The second term converges to σ^2 in probability. Now an application of Slutsky's Theorem gives that the LHS of (*) converges to a $\sigma^2(\chi_1^2 - 1)$ distribution. Thus we need n as the normalizing factor to get a non-degenerate limiting distribution. However, the limiting distribution under degeneracy is no longer normal.

Nonparametric Inference: Module 6[1]

What we provide in this module

- Functional in two sample problems
- Estimability and kernels
- Symmetry of kernels
- Two sample U statistic: Definition & properties
- Examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Functional

Suppose $X_1, X_2, ..., X_n$ are iid observations from F and $Y_1, Y_2, ..., Y_m$ are iid observations from G. The two samples are independent. F and G are unknown but known to be continuous. Suppose $\mathcal{F}$ is the class of all absolutely continuous DFs. Then $\theta(F, G)$ is a functional defined on $\mathcal{F} X \mathcal{F}$

2 Estimability & Kernels

$\theta(F, G)$ is said to be estimable if there exists a $\phi(X_1, X_2, ..., X_a, Y_1, Y_2, .., Y_b)$,$a, b \leq n$, such that,

$$E_F\{\phi(X_1, X_2, ..., X_a, Y_1, Y_2, .., Y_b)\} = \theta(F, G) \text{ for all } (F, G) \in \mathcal{F} X \mathcal{F}.$$

ϕ is called a kernel for estimation of $\theta(F, G)$. Minimum value of (a, b) with $a, b \leq n$, ensuring the above is called the degree of the kernel.

3 Symmetric Kernels

A kernel $\phi(X_{i_1}, X_{i_2}, ..., X_{i_a}, Y_{j_1}, Y_{j_2}, .., Y_{j_b})$, is said to be symmetric, if it remains unchanged for separate permutations of $\{i_1, i_2, ..., i_a\}$ and $\{j_1, j_2, ..., j_b\}$. As in the single sample case, corresponding to any asymmetric kernel there is always a symmetrized version. We are not going to give the rigorous proof and try to increase understanding by means of few examples.

3.1 Examples

Assume two independent samples $X_i \sim F, i = 1, 2, .., n$ and $Y_j \sim G, j = 1, 2, .., m$.

Example 1: Consider $\theta(F, G) = P(X_1 < Y_1)$. Since $\theta(F, G) = EI(X_1 < Y_1)$, we consider the kernel $\phi(X_1, Y_1) = I(X_1 < Y_1)$, where $I(\ )$ is the indicator function. Clearly $\phi(X_1, Y_1)$ is

a kernel of degree (1,1). The symmetry is immediate.

Example 2: Consider $\theta(F,G) = P(X_1 < Y_1, X_2 < Y_1)$. Since $\theta(F,G) = EI(X_1 < Y_1, X_2 < Y_1)$, we consider the kernel $\phi(X_1, X_2, Y_1) = I(X_1 < Y_1, X_2 < Y_1)$, where $I(\ )$ is the indicator function. Thus ϕ is a kernel of degree (2,1). Naturally ϕ does not change if X_1 and X_2 are interchanged. Hence ϕ is symmetric.

Example 3: Consider $\theta(F,G) = P(X_1 < Y_1 < X_2)$. Since $\theta(F,G) = EI(X_1 < Y_1 < X_2)$, we consider the kernel $\phi(X_1, X_2, Y_1) = I(X_1 < Y_1 < X_2)$, where $I(\ )$ is the indicator function. ϕ is a kernel of degree (2,1). But ϕ changes if X_1 and X_2 are interchanged. Thus ϕ is asymmetric. Corresponding symmetric kernel is $\frac{I(X_1<Y_1<X_2)+I(X_2<Y_1<X_1)}{2}$.

Example 4: For $\theta(F,G) = V_F(X_1) - V_G(Y_1)$, we consider the kernel $\phi(X_1, X_2, Y_1, Y_2) = \{X_1^2 - X_1X_2\} + \{Y_1^2 - Y_1Y_2\}$. ϕ is a kernel of degree (2,2). But ϕ changes if X_1 and X_2 and/ or Y_1 and Y_2 are interchanged. Thus ϕ is asymmetric and the corresponding symmetrized version can be obtained as $\frac{(X_1-X_2)^2-(Y_1-Y_2)^2}{2}$

4 U statistic-Definition

Define $C_1 = \{i_1, i_2, .., i_a : 1 \leq i_1 < i_2 < .. < i_a \leq n\}$ and $C_2 = \{j_1, j_2, .., j_b : 1 \leq j_1 < j_2 < .. < j_b \leq m\}$. Then we have in total $\binom{n}{a}\binom{m}{b}$ possible kernels of degree (a,b). If ϕ is asymmetric kernel of degree (a,b), then the corresponding U statistic is defined as

$$U = \{\binom{n}{a}\binom{m}{b}\}^{-1} \sum_{C_1} \sum_{C_2} \phi(X_{i_1}, X_{i_2}, ..., X_{i_a}, Y_{j_1}, Y_{j_2}, .., Y_{j_b}).$$

4.1 U statistic-Properties

As earlier $E(U) = \theta(F, G)$, that is, U is unbiased. Again U is a symmetric function of the observations. Naturally, U can be expressed in terms of the full set of order statistics. Since the full set of order statistics is complete sufficient under fairly general assumptions, U is an unbiased estimator based on the complete sufficient statistics. Thus U is MVUE of $\theta(F, G)$ even in two samples.

For the calculation of exact variance, define $\phi_{cd}(x_1, .., x_c, y_1, .., y_d) = E\{\phi(X_1, X_2, ..., X_a, Y_1, Y_2, .., Y_b)|$ $x_i, i = 1, 2, .., c, Y_j = y_j, j = 1, 2, .., d\}$

σ^2_{cd}= Covarince between $\phi(X_1.., X_c, X_{c+1}..X_a; Y_1, Y_2, .., Y_d, Y_{d+1}.., Y_b)$ and $\phi(X_1.., X_c, X'_{c+1}..X'_a; Y_1, Y_2,$ for $c + d > 1$, where X'_i are iid as F and Y'_j are iid as G. Then the exact variance of U can be expressed as $Var(U) = \sum\sum_{c+d\geq 1} \frac{\binom{a}{c}\binom{n-a}{a-c}}{\binom{n}{a}} \frac{\binom{b}{d}\binom{m-b}{b-d}}{\binom{m}{b}} \sigma^2_{cd}$

4.2 Examples

Assume X'_i are iid as F and Y'_j are iid as G.

Example 1: $\theta(F, G) = E(X_1) - E(Y_1)$. Take $\phi(X_1, Y_1) = X_1 - Y_1$, then ϕ is symmetric with degree $(1, 1)$. Then the corresponding U statistic is:

$$U = \frac{1}{\binom{n}{1}\binom{m}{1}} \sum_{1\leq i\leq n} \sum_{1\leq j\leq m} \phi(X_i, Y_j) = \bar{X} - \bar{Y}.$$

Thus we get the usual unbiased estimator. Since ϕ has degree (1,1), we have the choices (0,1),(1,0) and (1,1) for (c, d). Thus for the calculation of variance, we need to calculate σ^2_{01},σ^2_{10} and σ^2_{11}. Now $\sigma^2_{01} = cov(\phi(X_1, Y_1), \phi(X'_1, Y_1)) = Var_G(Y_1)$ $\sigma^2_{10} = cov(\phi(X_1, Y_1), \phi(X_1, Y'_1)) =$ $Var_F(X_1)$ $\sigma^2_{11} = cov(\phi(X_1, Y_1), \phi(X_1, Y_1)) = Var_F(X_1) + Var_G(Y_1)$ Then, a simple manipulation expresses the variance as

$$\begin{aligned} Var(U) &= \frac{\{V_F(X_1)+V_G(Y_1)+(m-1)V_F(X_1)+(n-1)V_G(Y_1)\}}{mn} \\ &= \frac{V_F(X_1)}{n} + \frac{V_G(Y_1)}{m} \end{aligned}$$

Example 2: Suppose $\theta = \theta(F, G) = P(X_1 < Y_1)$. We take $\phi(X_1, Y_1) = I(X_1 < Y_1)$, a symmetric kernel with degree $(1, 1)$. Corresponding U statistic is:

$$U = \tfrac{1}{mn} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} I(X_i < Y_j).$$

U is the famous Mann-Whitney statistic for comparing distributions. As earlier, we have the choices (0,1),(1,0) and (1,1) for (c, d). Thus for the calculation of variance, we need to calculate only σ_{01}^2,σ_{10}^2 and σ_{11}^2. Since $E\phi(X_i, Y_j) = \theta(F, G)$, $\sigma_{10}^2 = P(X_1 < Y_1, X_1 < Y_1^{'}) - \theta^2$. A simple application of conditioning methods give $\sigma_{10}^2 = \int (1 - G(x))^2 dF(x) - \theta^2$ Similar applications give:

$$\sigma_{01}^2 = P(X_1 < Y_1, X_1^{'} < Y_1) - \theta^2 = \int F(x)^2 dG(x) - \theta^2$$
$$\sigma_{11}^2 = Var(I(X_1 < Y_1)) = \theta(1 - \theta)$$

Using these, the final expression of the variance can be obtained.

Example 3: For $\theta(F, G) = Var(X_1) - Var(Y_1)$, take $\frac{(X_1-X_2)^2-(Y_1-Y_2)^2}{2}$, a symmetric kernel with degree $(2, 2)$. Corresponding U statistic simplifies to:

$$U = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (X_i - \bar{X})^2 - \tfrac{1}{m} \sum_{j=1}^{m} (Y_j - \bar{Y})^2 = s_X^2 - s_Y^2$$

Since ϕ has degree (2,2), we have a total of 3^2-1 possible choices, namely, (0,1),(1,0),(1,1),(1,2),(2,1),(and(2,2) and for (c, d). Then $\sigma_{01}^2 = cov(\phi(X_1, X_2, Y_1, Y_2), \phi(X_1^{'}, X_2^{'}, Y_1^{'}, Y_2))$ A simple calculation gives

$\sigma_{01}^2 = Var\{(Y_2 - \mu_Y)^2\}$, and $\sigma_{10}^2 = Var\{(X_1 - \mu_X)^2\}$. In a similar fashion the other terms and hence the final expression can be obtained. Although the derivation is straightforward but is time consuming. Now it is easy to observe that U is a difference of two univariate U statistics. Also the two univariate U statistics are independent. That is $U = U_1 - U_2$, where U_1 and U_2 are independent univariate U statistics. Then $Var(U) = Var(U_1) + Var(U_2)$. From the previous deductions, variances of U_1 and U_2 and hence the final expression of $Var(U)$ can be obtained.

5 Consistency of U statistic

Let U be the U statistic corresponding to a symmetric kernel of degree (a, b). Assume n observations are taken from F and m observations are taken from G. Suppose the following conditions hold:

(A1) $\sigma^2_{cd} > 0$ for $(c, d) = (0, 1), (1, 0), (a, a)$ and

(A2) Suppose for each $N = n + m$, there exists $n = n(N)$ and $m = m(N)$ such that as $N \to \infty$,

$\min(m, n) \to \infty$ and $\frac{n}{N}(\frac{m}{N}) \to \lambda(1 - \lambda)$ with $\lambda \in [0, 1]$.

Then under (A1) and (A2),

$Var(\sqrt{N}U) \to \sigma^2 = \frac{a^2\sigma^2_{10}}{\lambda} + \frac{b^2\sigma^2_{01}}{1-\lambda}$ Now we have $E(U) = \theta(F, G)$ and consider the representation

$Var(U) = \frac{1}{N}Var(\sqrt{N}U)$ Then considering the previous results, we observe that as $\min(m, n) \to \infty$, $Var(\sqrt{N}U) \to \sigma^2$ and $\frac{1}{N} \to 0$ Combining, we get that as $Var(U) \to 0$. Therefore, the consistency Of U statistic follows for two samples.

6 Asymptotic Distribution of U statistic

Result: Under conditions (A1) and (A2) above, as $N \to \infty$,

$$\sqrt{N}(U - \theta(F, G)) \xrightarrow{D} N(0, \sigma^2),$$

where $\sigma^2 = \frac{a^2\sigma^2_{10}}{\lambda} + \frac{b^2\sigma^2_{01}}{1-\lambda}$

Proof: Define, $Z_N = \sqrt{N}(U - \theta) - \frac{a\sqrt{N}}{n}\sum_{i=1}^{n}(\phi_{10}(X_i - \theta) - \frac{b\sqrt{N}}{m}\sum_{j=1}^{m}(\phi_{01}(Y_j - \theta)$. Then, it can be observed that

(i) $E(Z_N) = 0$

(ii) By CLT, the second and third terms in the RHS of Z_N are asymptotically independent and normal

(iii) $E(Z_N^2) \to 0$ as $N \to \infty$.

Thus

$$Z_N \xrightarrow{P} 0$$

as $N \to \infty$ and hence

$$\sqrt{N}(U - \theta)$$

and

$$\frac{a\sqrt{N}}{n}\sum_{i=1}^{n}(\phi_{10}(X_i) - \theta) + \frac{b\sqrt{N}}{m}\sum_{j=1}^{m}(\phi_{01}(Y_j) - \theta)$$

have the same asymptotic distribution. The latter is asymptotically $N(0, \sigma^2)$ and hence the proof is complete.

6.1 Examples

Example 1 revisited: For $\theta(F, G) = E(X_1) - E(Y_1)$ applying the above result and the expressions of σ_{10}^2 and σ_{01}^2, we get as $N \to \infty$, $\sqrt{N}(\bar{X} - \bar{Y} - \theta) \xrightarrow{D} N(0, \sigma^2)$, where $\sigma^2 = \frac{\sigma_{10}^2}{\lambda} + \frac{\sigma_{01}^2}{1-\lambda} = \frac{V_F(X_1)}{\lambda} + \frac{V_G(Y_1)}{1-\lambda}$.

Example 2 revisited: Suppose $\theta = \theta(F, G) = P(X_1 < Y_1)$. Then as $N \to \infty$, $\sqrt{N}(U - \theta) \xrightarrow{D} N(0, \sigma^2)$, where

$$\sigma^2 \quad = \frac{\int (1-G(x))^2 dF(x) - \theta^2}{\lambda} + \frac{\int F(x)^2 dG(x) - \theta^2}{1-\lambda}.$$

If $F = G$, then $\theta = \frac{1}{2}$, $\sigma_{10}^2 = \sigma_{01}^2 = \frac{1}{12}$. Thus under $F = G$, we have

$$\sqrt{N}(U - \frac{1}{2}) \xrightarrow{D} N(0, \frac{1}{12\lambda(1-\lambda)})$$

Example 3 revisited: For $\theta(F, G) = Var(X_1) - Var(Y_1)$, we get with the already derived

expressions, as $N \to \infty$, $\sqrt{N}(U - \theta) \xrightarrow{D} N(0, \sigma^2)$, where

$$\sigma^2 \quad = \tfrac{Var\{(X_1-\mu_X)^2\}}{\lambda} + \tfrac{Var\{(Y_2-\mu_Y)^2\}}{1-\lambda}.$$

It is easy to observe that the asymptotic variance is the weighted sum of the asymptotic variances of two univariate U statistics.

Nonparametric Inference: Module 7[1]

What we provide in this module

- Nonparametric hypothesis testing and confidence interval: Why the need?
- Components of a nonparametric test
- Consistency of tests-Some basics
- Different nonparametric hypothesis testing problems
- A distribution free confidence interval for the population quantile

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Nonparametric hypothesis testing and confidence interval: Why the need?

The central idea of parametric (or classical) inference is the assumption regarding the underlying population. The entire theory of the parametric inference is developed under this assumption and consequently these procedures are valid as long as these assumptions are satisfied. For example, students t test is appropriate only when the underlying distribution is normal. But normal is not the only distribution having applications in real life. For example, in survival trials, the lifetime distributions are mostly exponential, gamma or Weibull, that is non normal. Therefore, inference based on t test in such a situation is often misleading. Therefore, parametric tests are useful only when the experimenter is sufficiently confident about the underlying distribution. But unfortunately, the available methods of identifying the underlying distribution is limited to standard distributions only.

Therefore, it will be better if hypothesis testing procedures can be developed with minimal assumptions about the underlying distribution(say, continuity of observations). Suppose there exists a statistic T, relevant to the testing problem such that the exact and/or conditional and/or asymptotic distribution of T under the null hypothesis is independent of the underlying distribution. Naturally, the significance level for the test based on T does not depend on the underlying distribution, that is, these tests are level robust. Tests based on T are commonly termed as nonparametric or distribution free.

In real practice, T is often formed by taking only the signs or ranks of the actual observations. But such a T does not take into account the full information contained in the individual observations and hence are often less efficient than their parametric counterparts. However, nonparametric tests are the valid choices as long as the validity of the parametric assumptions are questionable.

As in parametric inference problems, the experimenter may be interested in constructing a confidence interval with coverage probability independent of the underlying distribution. However, in nonparametric inference, the unknown quantity of interest are mostly location

or scale parameters. As in the parametric counterpart, we can invert the acceptance region of a distribution free test to get a distribution free confidence interval with high coverage probability. These will be discussed later. However, if the unknown quantity of interest is a population quantile, we can develop a distribution free confidence interval based on the sample order statistics.

2 Components of a nonparametric test

In most of the problems of nonparametric inference, the underlying distribution is not specified except for continuity. Therefore the available methods(e.g. Neyman Pearson lemma, likelihood ratio method) of test construction are of no use and hence, we need to develop tests with intuitive appeal. In particular, if we can identify a distribution free statistic T for the problem, we can develop a distribution free test. However, for a meaningful development, we maintain the following sequence:

1. Model assumption(i.e. the minimal set of assumptions about the underlying distribution)
2. Hypothesis of interest(i.e. specification of the null and alternative hypotheses)
3. Available tests for the problem(i.e. the usual parametric procedures for specific probability models)
4. Suggesting a distribution free statistic(i.e. specifying some T)
5. Justifying the form of critical region.
6. Investigating unbiasedness and consistency of the suggested test and finally
7. Providing a large sample test corresponding to the given test.

3 Consistency of tests-Some basics

However, tests based on such a T are often far from being optimal and hence properties like maxmising power is not immediate. Consistency is, therefore, an important concern to measure the sensitivity of the test in large samples to a little departure from the null hypothesis. We provide below the notion of consistency of tests in little details.

Suppose $X_1, X_2, .., X_N$ are iid observations from an unknown distribution G. We are interested in testing $H_0 : G \in \Omega_0$ against $H_a : G \in \Omega_a$, where $\Omega_0(\Omega_a)$ is the class of distributions specified by the null(alternative) hypothesis.

A sequence of tests $\{\phi_N\}$ is said to be consistent if for every $G \in \Omega_a$,

$$E_G \phi_N \to 1$$

as $N \to \infty$.

However, such a definition is not easy to check and hence we require some sufficient conditions.

Assume that ϕ_N is a right tailed test based on a statistic T_N. Then $\{\phi_N\}$ is consistent if

(i) ϕ_N is asymptotically size α, that is, $E_G \phi_N \to \alpha$ for every $G \in \Omega_0$ and

(ii) $T_N \xrightarrow{P} \mu(G)$, with $\mu(G)$ satisfying

$$\begin{aligned} \mu(G) = & \ \mu_0 \quad \text{for all } G \in \Omega_0 \\ > & \ \mu_0 \quad \text{for all } G \in \Omega_a. \end{aligned}$$

The requirement on asymptotic size is not immediate in practice. However, if we can establish asymptotic normality of T_N for all $G \in \Omega_0$, then the requirements in (i) and (ii) above are automatically satisfied. Therefore, if there exists a constant $\sigma_0(> 0)$ such that as $N \to \infty$

$$\frac{\sqrt{N}(T_N - \mu_0)}{\sigma_0} \xrightarrow{D} N(0,1) \quad \text{for all } G \in \Omega_0,$$

then ϕ_N is consistent. Therefore asymptotic normality is a stronger condition than those required in consistency.

However, asymptotic normality or asymptotic size condition is not immediate and hence we provide some simpler conditions. Assume that G is indexed by a real parameter θ, that is, $G(x) = G(x, \theta)$ and the testing problem can be expressed as $H_0 : \theta = \theta_0$ against $H_a \theta > \theta_0$. Suppose a level α test rejects the null hypothesis if $S_N(\theta_0) \geq c_N(\theta_0)$, where $S_N(\theta_0)$ is so constructed that

$$S_N(\theta_0) \xrightarrow{P} \Delta(\theta, \theta_0)$$

with

$$\begin{aligned} \Delta(\theta, \theta_0) = & \quad 0 \quad \text{for } \theta = \theta_0 \\ & > 0 \quad \text{for all } \theta > \theta_0 \end{aligned}$$

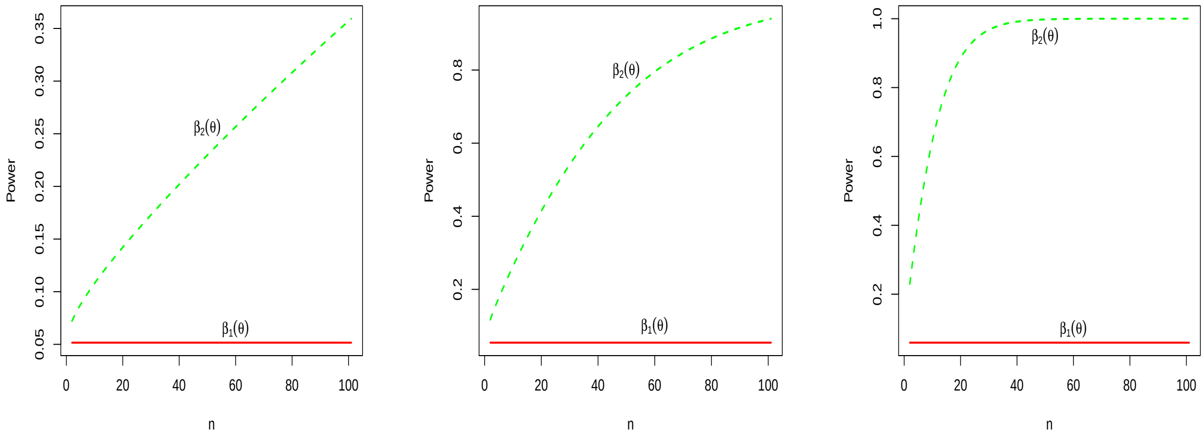
and $\lim_{N \to \infty} c_N(\theta_0) \leq 0$. Then the test based on $S_N(\theta_0)$ is consistent.

However lack of consistency is a serious problem. We shall explain such a consequence through an example. Suppose $X_1, .., X_n$ are iid observations from a Cauchy distribution with location parameter θ. Suppose we want to test $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$. We suggest two possible tests, one based on sample mean $\bar{X}$ and the another based on the sample median $\tilde{X}$.

Test 1: Reject H_0 if $\bar{X} > c$, where c is such that $P_{H_0}(\bar{X} - \theta_0 > c) = .05$. Since under the null hypothesis $\bar{X} - \theta_0$ has a Cauchy(0,1) distribution, we get $c = tan(.45\pi) = 6.314$. Then the power function is $\beta_1(\theta) = P_\theta(W > \theta_0 - \theta + 6.314)$, where $W \sim Cauchy(0, 1)$.

Test 2: Reject H_0 if $\tilde{X} > c'$, where c' is such that $P_{H_0}(\tilde{X} - \theta_0 > c') = .05$. But the distribution of $\tilde{X} - \theta_0$ is not easy to find. However, it is known that $\sqrt{n}(\tilde{X} - \theta)$ is asymptotically $N(0, \pi^2/4)$ distribution under any θ. Then considering the asymptotic distribution, we set $c' = \tau_{.05} \frac{\pi}{2\sqrt{n}} = \frac{2.584}{\sqrt{n}}$. Then the power function becomes $\beta_2(\theta) = P_\theta(\tilde{X} - \theta > \theta_0 + \frac{2.584}{\sqrt{n}})$. Now it is easy to observe that both the tests are asymptotically size .05, but the latter is based on an asymptotically normal statistic whereas the former involves an inconsistent statistic. For better visualisation of the effect of consistency on power, we provide a plot of both the power functions taking $\theta_0 = 0$. For a better visualisation, we compute powers for three choices of θ and plot the power for various values of the sample size(n). All these can be

found in Figure 1 below. We shall explain the difference in the nature of power functions



(a) $\theta = .20$ (b) $\theta = .50$ (c) $\theta = 1.0$

Figure 1: Nature of power for varying sample size(n)

for varying n. Suppose we fix $\theta_0 = 0$ and $\theta = .2$. It is easy to observe that the power for Test 1 remains very close to .05 for varying n whereas the power for Test 2 increases sharply. The same is observed for the assumed choices of θ. However, the rate increase for the power of Test 2 is higher for higher values of θ. This is expected, as the first test is based on an inconsistent estimator(i.e. $\bar{X}$), i.e. , it does not concentrate around the true value for large n. Consequently the power increases at a very slow rate for Test 1. On the other hand, $\tilde{X}$ is consistent and hence for large n approaches the true parameter and consequently power increases for increasing values of n.

4 Different nonparametric hypothesis testing problems

Now we shall discuss the different types of hypotheses considered in nonparametric hypothesis testing together with their relevances. Based on the availability of data, hypothesis testing problems are either single sample or two sample or multi-sample. We, therefore, discuss hypotheses for each type of problems.

4.1 Single sample problems

Suppose $X_1, X_2, .., X_n$ are iid observations from an unknown distribution F. F is unknown but known to be continuous. Then depending on the requirement, we have the following different hypotheses.

4.1.1 Problem of location

Suppose $p \in (0,1)$ is a known quantity and let $\theta(F) = \xi_p(F)$ be the quantile of order p for F, that is $F(\theta(F)) = p$. Then the problem of location is to test $H_0 : \theta(F) = \theta_0$ against one of the alternatives $H_a : \theta(F) > \theta_0$ or $H_a : \theta(F) < \theta_0$ or $H_a : \theta(F) \neq \theta_0$ for some known θ_0. For the above hypothesis, only the continuity of F is required. However, if we consider the same hypothesis with the added assumption of symmetry of F, it will be the problem location under symmetry. Since, the main concern in a problem of location or location under symmetry, is the location(e.g. median), the hypothesis is an analogue to the test of a location parameter in parametric counterpart.

4.1.2 Goodness of fit problem

In a goodness of fit problem, the interest lies in investigating whether the sample comes from a specified distribution. Then the hypothesis of interest in a goodness of fit problem can be described as $H_0 : F(x) = F_0(x)$ for all x, where F_0 is a completely known DF. The alternative hypothesis is naturally $H_a : F(x) \neq F_0(x)$ for at least one x. The problem of goodness of a fit also arises in parametric inference after some known distribution is fitted to a data.

4.2 Two sample problems

Suppose $X_i, i = 1, 2, .., n$ and $Y_j, j = 1, 2, .., m$ are independent samples from unknown distributions F and G respectively. We only assume the continuity of observations from F and G. The basic hypothesis in any two sample problem is $H_0 : F(x) = G(x)$ for all x

against the usual one sided or two sided hypothesis. Defining $\Gamma_0 = \{(F,G) : F(x) = G(x) \forall x\}$, we can express the null hypothesis as $H_0 : (F,G) \in \Gamma_0$. Depending on different specifications of Γ_0 and $\Gamma_a = \{(F,G) : F(x) \neq G(x)$ for some x $\}$, we have the following possible hypotheses.

1. **General/ Homogeneity Alternative:** Suppose the two underlying populations differ in any manner(in location, scale or skewness). Then such an alternative hypothesis can be expressed as $\Gamma_a = \{(F,G) : F(x) \neq G(x) \text{ for some x}\}$. The general hypothesis is also termed as Homogeneity alternative.

2. **Stochastic Alternative:** A stochastic alternative is a restricted alternative, where

$$\Gamma_a = \{(F,G) : G(x) \geq F(x) \text{ for all x with strict inequality for some x}\}.$$

Actually $G(x) \geq F(x)$ for all x implies that X observations are tend to be larger than Y observations or, in other words, X is *stochastically larger* than Y. This is a general class of alternatives.

3. **Location Alternative:** Suppose $G(x) = F(x - \theta)$, where $\theta \neq 0$. That is, the underlying distributions differ only in location. Then $F(x) > G(x)$ or $F(x) = G(x)$ or $F(x) < G(x)$ according as $\theta < 0$ or $\theta = 0$ or $\theta > 0$. Thus the null hypothesis can be restated as $H_0 : \theta = 0$ and the alternative is either $H_a : \theta > 0$ or $H_a : \theta < 0$ or $H_a : \theta \neq 0$. Clearly $\theta > 0$ implies $\Gamma_a = \{(F,G) : \text{ F is shifted to the right of G for some x}\}$. This is a special case of a stochastic alternative, where $G(x) = F(x - \theta)$. Stochastic alternative,in general, relates to the location alternative in a less restrictive sense because $G(x) > F(x)$ indicates larger Y observations and hence corresponds to larger location of Y observations.

4. **Scale Alternative:** Suppose $G(x) = F(\frac{x}{\sigma})$ with $\sigma > 0$, that is, the two underlying populations are assumed to differ only in scale. Since, $F(x) \gtreqless G(x) \Leftrightarrow \sigma \gtreqless 1$, the null hypothesis reduces to $H_0 : \sigma = 1$. The alternative is either of $H_a : \sigma > 1$ or $H_a : \sigma < 1$

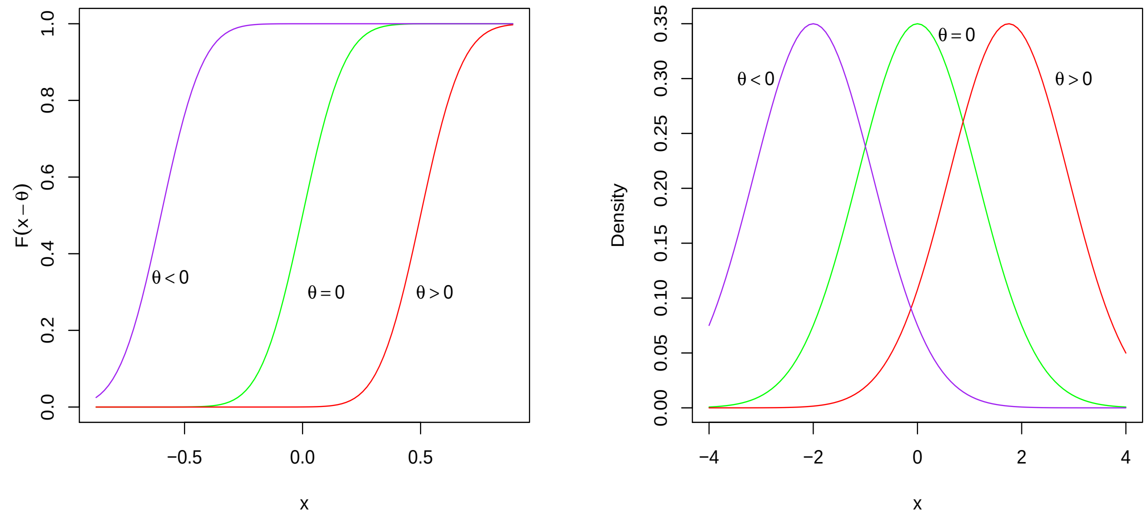


Figure 2: Effect of location on the distribution function and density

or $H_a : \sigma \neq 1$. Scale alternative can also be viewed as a stochastic alternative, where $G(x) = F(\frac{x}{\sigma})$.

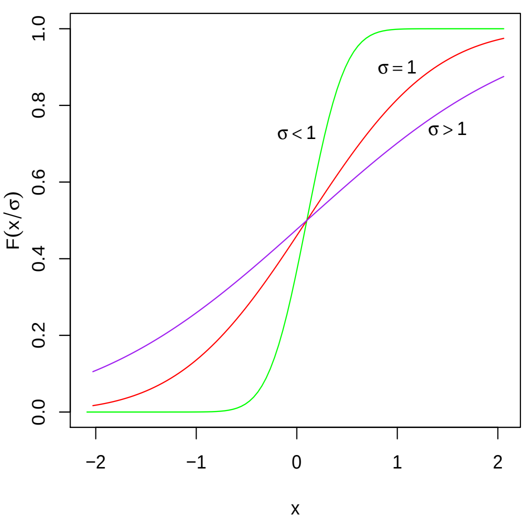
It is worthwhile to mention that tests meant for general or stochastic alternative can also be used for location and scale alternatives. But they will be less efficient than the tests developed for the specific alternative. A similar set of alternatives can be found corresponding to any multiple sample problems and will be discussed later considering specific situations and hence are not discussed separately.

4.3 Paired sample problems

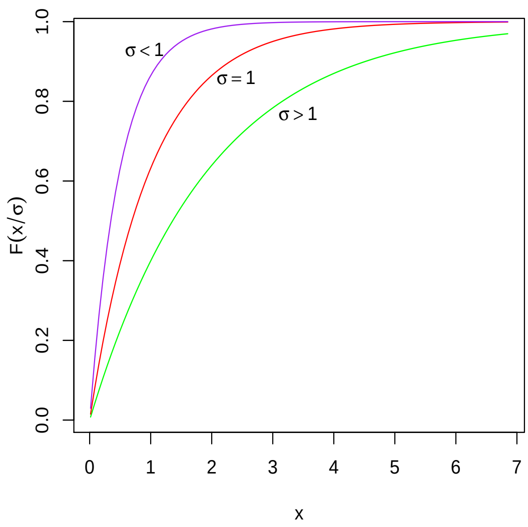
Suppose $(X_i, Y_i) i = 1, 2, .., n$ are sample observations from an unknown bivariate distribution $F(x, y)$ with $F_X(x)$ and $F_Y(y)$ as the marginal DF's. Then the following two hypotheses are of main interest:

1. **Problem of association:** Here the interest lies in testing

$$H_0 : F(x, y) = F_X(x)F_Y(y) \text{ for all (x,y)}$$



(a) Symmetric population



(b) Asymmetric population

Figure 3: Effect of varying scale on distribution functions

against the alternative

$$H_a : F(x, y) \neq F_X(x)F_Y(y) \text{ for some (x,y).}$$

2. **Problem of location:** Suppose X represents the response before a drug is administered and Y denote that after the application of the drug. Then naturally Y are influenced by X and hence X and Y are correlated. Then the natural objective in this situation is to determine whether the drug has any effect. In statistical terms, this means X and Y are exchangeable. Thus the null hypothesis can be expressed as

$$\begin{aligned} H_0 : & \quad \text{X and Y are exchangeable.} \\ \equiv & \quad H_0 : (X, Y) \stackrel{D}{=} (Y, X) \\ \equiv & \quad H_0 : F(x, y) = F(y, x) \ \forall \ (x, y). \end{aligned}$$

Define $D = Y - X$, Then under the null hypothesis distribution of D is symmetric about the origin. If X and Y differs only in location under the alternative, then

$D - \Delta \stackrel{D}{=} -(D - \Delta)$, where Δ is the location difference. Naturally, under the null hypothesis, the distribution of D has median at the origin but under the alternative, the median becomes Δ. Then the problem reduces to testing $H_0 : \Delta = 0$ against all alternatives. However, the median of the distribution of the difference is not always the difference of the the marginal medians. If the marginal distributions and the distribution of the difference are all symmetric, then median of the distribution of difference and the difference of the two medians coincide(see, Gibbons and Chakraborti, 2006, for details).

5 Distribution free confidence interval for quantiles

Suppose $X_i, i = 1, 2, .., n$ are iid observations from a continuous but unknown DF $F(x)$. The objective is to provide a confidence interval of the p th order quantile ξ_p satisfying $F(\xi_p) = p$. Since ξ_p is a population quantile , it is natural to consider intervals based on the sample quantiles as a confidence interval. Thus we can start with $X_{(r)}$ (i.e. the sample $\frac{r}{n}$ th quantile) and $X_{(s)}$ (i.e. the sample $\frac{s}{n}$ th quantile). That is, we suggest to consider the interval $(X_{(r)}, X_{(s)})$ with $r < s$ as a confidence interval for ξ_p. Now we shall show that the coverage probability of $[X_{(r)}, X_{(s)}]$ is independent of any F. Note that for any k, $X_{(k)} \leq \xi_p \Leftrightarrow Z \geq k$, where Z has a binomial distribution with parameters n and success probability $F(\xi_p) = p$. Thus

$$
\begin{aligned}
P\{[X_{(r)}, X_{(s)}] \ni \xi_p\} &= P(X_{(r)} \leq \xi_p) - P(X_{(s)} \leq \xi_p) \\
&= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} = \gamma(n, r, s)
\end{aligned}
$$

Thus $[X_{(r)}, X_{(s)}]$ is a confidence interval for ξ_p with confidence coefficient $\gamma(n, r, s)$. Clearly, $\gamma(n, r, s)$ is independent of any F and hence $[X_{(r)}, X_{(s)}]$ gives a distribution free confidence interval of ξ_p. However, in practice the confidence coefficient is set at least $(1 - \alpha)$ with

specified α and hence we choose s and r in such a way that $\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \geq 1 - \alpha$ is satisfied.

But such a choice is not unique as we are determining r and s from a single equation. However, if we want to get a confidence interval with smallest length, we can determine r and s such that $E(X_{(s)} - X_{(r)})$ is a minimum for a given α. The most common approach to achieve uniqueness is to assign equal probability at each tail. That is, for a given α, we determine r and s satisfying $\sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\alpha}{2}$ and $\sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\alpha}{2}$. As another alternative $r + s$ is taken as $n + 1$ so that the confidence limits become equidistant from the two endpoints.

Nonparametric Inference: Module 8[1]

What we provide in this module

- A motivating example explaining the need of alternative hypothesis testing procedure
- Sign Test: Definition and Properties
- Intuitive justification of the form
- Optimality of Sign test
- Consistency of Sign test

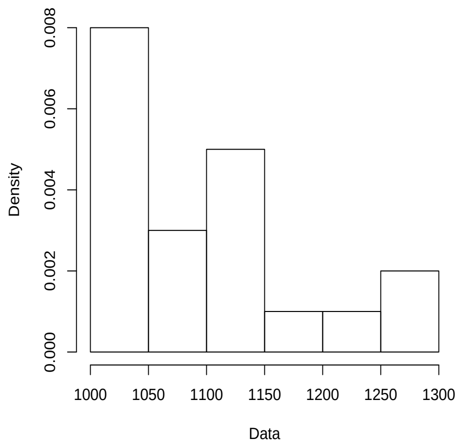
[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A motivating example

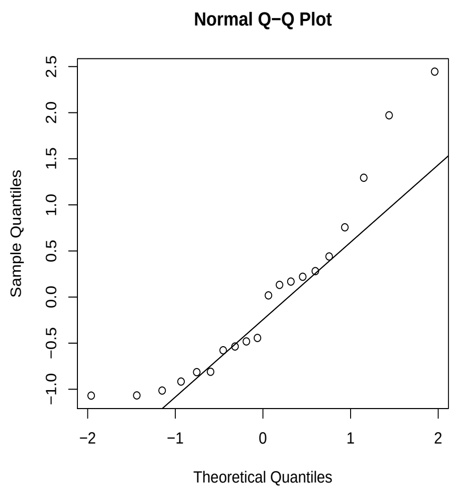
We start with a motivating example. Consider the following data on the lifetime of an electric equipment:

1046.541 1110.841 1259.690 1014.233 1156.425 1001.439 1299.962 1116.045
1022.895 1106.415 1023.236 1093.674 1103.354 1005.930 1202.124 1001.251
1129.546 1051.215 1043.066 1054.430.

Suppose interest is to test the null hypothesis that the mean lifetime is 1050 hours. The usual practice is to use Student's t distribution for the purpose. But the question is natural " Whether normality assumption holds?" We perform some exploratory data analysis. We provide below the histogram and Q-Q plot for the data.



(a) The Histogram



(b) Q-Q plot

Figure 1: Histogram and Q-Q plot of the data

The histogram shows that the distribution of the data is far from symmetric. The QQ plot clearly reveals the non normality of data. Then t test is not appropriate.A test for this data will be appropriate , if the assumed distribution is appropriate. However, deciding an

appropriate distribution suffers from subjectivity and no such thumb-rule is present. Thus, we need alternative procedures to test the hypothesis appropriately.

2 What is Sign Test?

This is a nonparametric analogue of Student's t test for the mean(i.e. a location parameter) of the population. The t-test is based on the assumption of normality of the underlying population. Sign test is a nonparametric alternative to t test. This test does not require the assumption of normality. It also provides a test of location but uses quantiles of the distribution as the location parameter. Moreover, sign test is based on only the continuity of the underlying population.

2.1 Assumptions & the hypothesis

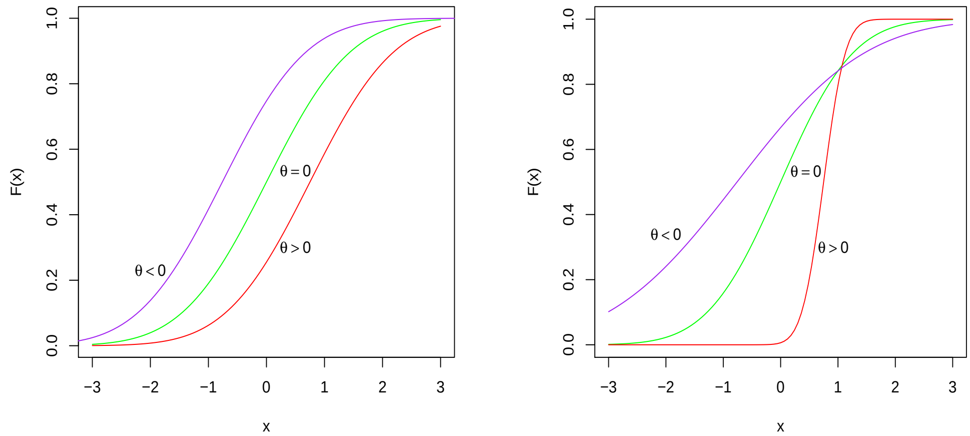
Suppose $X_1, X_2, .., X_n$ are iid observations from a population characterised by the DF F, where F is unknown but assumed to be continuous. Suppose $\theta(F)$ is the quantile of order p, that is $F(\theta(F)) = p$ for known p. Then in Sign test, the objective is to test $H_0 : \theta(F) = \theta_0$ against one of the alternatives $H_a : \theta(F) > \theta_0$ or $H_a : \theta(F) < \theta_0$ or $H_a : \theta(F) \neq \theta_0$ for some known θ_0. For our discussion, we choose $p = 0.5$ so that $\theta(F)$ reduces to the median.

2.2 Sign test statistic-The intuitive argument

Note that if the observed data is consistent with $\theta(F) = \theta_0$, then one can expect that almost 50% values of the data set lie above and below θ_0. This suggests to use the number of observations exceeding θ_0 as the test statistic. Formally this implies the use of the statistic $S(\theta_0) = \sum_{i=1}^{n} I(X_i - \theta_0 > 0)$ Since S counts the number of positive signs among $X_i - \theta_0, i = 1, 2, .., n$, the test based on S is called Sign test.

2.3 Sign test statistic: Another look

Assume that $\theta_0 = 0$. Then $\theta(F) > 0 \Leftrightarrow F(0) < .5$, that is $P(X_1 > 0) > .5$. Similarly, $\theta(F) < 0 \Leftrightarrow F(0) > .5$, that is $P(X_1 < 0) > .5$. Thus $\theta(F) > 0(or < 0)$ implies more positive(negative) observations. The following figures will make the idea clear.



(a) Equal scale

(b) Varying scale

Figure 2: Location of θ

Consider the alternative $H_0 : \theta > 0$. This suggests to use the number of positive observations as our test statistic. Formally this implies the use of the statistic $S = \sum_{i=1}^{n} I(X_i > 0)$. Similarly, the form of the statistic for other alternatives can also be justified.

3 Distribution of S

Assume that $P(X_i = \theta_0) = 0$ for every $i = 1, 2, .., n$. Since each $I(X_i - \theta_0 > 0)$ can be thought of a Bernoulli random variable, S can be looked upon as a sum of n Bernoulli random variables. Since observations are iid, $I(X_i - \theta_0 > 0)$ are iid random variables. Now the distribution of each $I(X_i - \theta_0 > 0)$ is Bernoulli with success probability $P(X_1 > \theta_0)$. Thus

S is the sum of n iid Bernoulli random variables with success probability $P(X_1 > \theta_0)$.

We see that, S has a $Binomial(n, P(X_1 > \theta_0)$ distribution. Naturally the success probability $P(X_1 > \theta_0)$ depends on the underlying F. However under the null hypothesis $F(\theta_0) = 0.5$ and hence S becomes a distribution free statistic.Therefore tests based on S are exactly nonparametric.

4 Critical region

Since $S \sim Binomial(n, \frac{1}{2})$, under H_0, $E(S) = \frac{n}{2}$. However, under any $\theta(F)$, $E(S) = n(1 - F(\theta_0))$. Suppose $\theta > \theta_0$, then it is expected to have more than 50% observations exceeding θ_0. Thus $S(\theta_0)$ is expected to be larger under $\theta > \theta_0$ than under $\theta = \theta_0$. Therefore, larger values of $S(\theta_0)$ indicates evidence against $\theta = \theta_0$.Naturally a right tailed test based on $S(\theta_0)$ seems appropriate for testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$.

Again less than 50% observations exceeding θ_0 are expected under $\theta < \theta_0$. Thus $S(\theta_0)$ is expected to be smaller under $\theta < \theta_0$ than under $\theta = \theta_0$. Thus a left tailed test based on $S(\theta_0)$ seems appropriate for the alternative $H_a : \theta < \theta_0$. However, if $\theta \neq \theta_0$, then $S(\theta_0)$ is expected to be either smaller or larger than under $\theta = \theta_0$. Therefore, a two tailed test based on $S(\theta_0)$ is appropriate for the alternative $H_a : \theta \neq \theta_0$.

5 Symmetry of S

Since the distribution of $S(\theta_0)$ is Binomial(n,.5) under the null hypothesis, $S(\theta_0)$ has a symmetric distribution about $\frac{n}{2}$. We explore the implication of the symmetric nature. Suppose we have observed $Y_i = 2\theta_0 - X_i$ for each i, instead of X_i. Under $\theta = \theta_0$, the median of the distributions of both Y_i and X_i is θ_0. Then tests for the two sided alternative based on Y_is and X_is are expected to give similar results. But test applied on Y's will give similar result to that applied on X's, if the statistic has a symmetric distribution. This gives the

justification of the requirement of symmetry.

5.1 Different Tests

Since, S has a discrete distribution, tests based on it will be randomized. For the alternative $H_a : \theta > \theta_0$, a size α test can be expressed as $\phi^0 = I(S > S_\alpha) + aI(S = S_\alpha)$, where S_α is such that $E_{H_0}\phi^0 = \alpha$.

For the alternative $H_a : \theta < \theta_0$, a size α test can be expressed as $\phi^0 = I(S < S_{1-\alpha}) + aI(S = S_{1-\alpha})$, where $S_{1-\alpha}$ is such that $E_{H_0}\phi^0 = \alpha$.

For the alternative $H_a : \theta \neq \theta_0$, since the distribution of S is symmetric about $\frac{n}{2}$ under the null hypothesis, a size α test can be expressed as $\phi^0 = I(|S-\frac{n}{2}| > S_{\frac{\alpha}{2}}) + aI(|S-\frac{n}{2}| = S_{\frac{\alpha}{2}})$ with $E_{H_0}\phi^0 = \alpha$.

5.2 Test based on p values

Suppose S_{obs} is the observed value of S. For the alternative $H_a : \theta > \theta_0$, the one sided p value is $P_{H_0}(S \geq S_{obs})$. We accept the null hypothesis if this p value exceeds α. For the alternative $H_a : \theta < \theta_0$, the one sided p value is $P_{H_0}(S \leq S_{obs})$. We accept the null hypothesis if this p value exceeds α. However, for the two sided alternative $H_a : \theta \neq \theta_0$, the two sided p value is $2min\{P_{H_0}(S \geq S_{obs}), P_{H_0}(S \leq S_{obs})\}$. We reject the null hypothesis if this p value does not exceed α.

6 Presence of ties

We have already assumed continuity of F so that $P(X_i = \theta_0) = 0$ for every $i = 1, 2, .., n$. But in practice, we can have observations equal to θ_0. Thus we get some zero's in S. Presence of a large number of 0's can give misleading results. The usual method, in this context

is to ignore the zero differences. However ties can occur theoretically if $P(X_1 = \theta_0) > 0$. Suppose, now we are interested in $H_0 : P(X_1 > \theta_0) = P(X_1 < \theta_0)$. Let $S^+(S^-)$ denote the number of positive(negative) signs in such a case. Then $S^+ + S^- < n$, and hence the conditional distribution of S^+ given $S^+ + S^-$ is again binomial with parameters $S^+ + S^-$ and $p = \frac{P(X_1>\theta_0)}{P(X_1>\theta_0)+P(X_1<\theta_0)}$. Thus $p = \frac{1}{2}$ under the null hypothesis and therefore, S^+) can be taken as our new statistic and tests based on it can be performed as earlier. These tests are known as conditional sign tests.

7 Optimality of Sign Test

Consider testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$. Define $\Gamma_0 = \{F : F(\theta_0) = \frac{1}{2}$ and $\Gamma_a = \{F : F(\theta_0) < \frac{1}{2}$. Then the above hypothesis testing can be equivalently expressed as testing $H_0 : F \in \Gamma_0$ against $H_a : F \in \Gamma_a$. Then H_0 and H_a are both composite. It can be shown that the UMP size α test for the above testing problem is nothing but the Sign test based on S. In a similar way the two sided Sign test is UMPU size α(see, Fraser, 1957, for details).

8 Consistency of Sign test

Consider testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$. Now Sign test can be equivalently expressed in terms of $\frac{S}{n}$. For simplicity assume $\theta_0 = 0$. Since S has a binomial distribution, we have under any $\theta = \theta(F)$,

$$\frac{S}{n} \xrightarrow{P} P(X_1 > \theta_0).$$

Being consistent with our development, we define $\mu(F) = P(X_1 > \theta_0)$. Then $\mu(0) = \frac{1}{2}$ is the value of $\mu(F)$ under the null hypothesis and hence

$$\begin{aligned} \mu(F) &= \frac{1}{2} \text{ if } \theta = \theta_0 \\ &> \frac{1}{2} \text{ if } \theta > \theta_0 \end{aligned}$$

Thus Sign test is consistent against the alternative $H_a : \theta > \theta_0$. Consistency against the other alternatives can be proved also.

Nonparametric Inference: Module 9[1]

What we provide in this module

- Unbiasedness of Sign Test
- Confidence interval estimation using Sign test
- Large sample test
- Sample size determination
- Applications of Sign Test

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Unbiasedness of Sign Test

Suppose $X_1, X_2, .., X_n$ are iid observations from a population characterised by the DF F, where F is unknown but assumed to be continuous. Suppose $\theta(F)$ is the quantile of order p, that is $F(\theta(F)) = p$ for known p. Consider testing $H_0 : \theta(F) = \theta_0$ against the alternative $H_a : \theta(F) > \theta_0$. Then the power function can be expressed as

$$\begin{aligned} E_\theta \phi^0 &= P_\theta(S(\theta_0) > S_\alpha) + aP_\theta(S(\theta_0) = S_\alpha) \\ &= aP_\theta(S(\theta_0) > S_\alpha - 1) + (1-a)P_\theta(S(\theta_0) > S_\alpha) \end{aligned}$$

Now observe that $S(\theta_0) = \sum_{i=1}^n I(X_i > \theta_0)$ is expected to be larger under θ than under θ_0. Thus $S(\theta_0)$ under θ is stochastically larger than that under θ_0. Then we find that $E_\theta \phi^0 \geq P_{\theta_0}(S(\theta_0) > S_\alpha) + aP_{\theta_0}(S(\theta_0) = S_\alpha)$. It is easy to observe that the RHS of above equals $E_{\theta_0}\phi^0 = \alpha$. Thus we find that $E_\theta \phi^0 \geq \alpha$ for any $\theta \geq \theta_0$ and hence unbiasedness follows.Unbiasedness for the other tests can also be proved in a similar way.

2 Confidence interval using Sign test

In parametric inference, we often obtain a confidence interval from the acceptance region of the test. The same technique, though difficult, can also be adopted in nonparametric procedures. Thus we can use the cut offs of two sided Sign test to get a confidence interval of θ with confidence coefficient at least $(1-\alpha)$. Consider the acceptance region of the non randomized level α Sign test for $H_0 : \theta = 0$ against $H_a : \theta \neq 0$. With the already introduced notations, the acceptance region can be expressed as $c \leq S(0) \leq n - c$, where c is such that $P_{H_0}(c \leq S(0) \leq n - c) \geq 1 - \alpha$. We have already seen that $S(\theta)$ is non increasing in θ. Then using the properties of order statistics, we at once obtain

$$S(\theta) \leq n - c \Leftrightarrow \theta \geq X_{(c)}$$

and

$$S(\theta) \geq c \Leftrightarrow \theta < X_{(n-c+1)}$$

. Thus $c \leq S(\theta) \leq n - c$ is equivalent to $X_{(c)} \leq \theta < X_{(n-c+1)}$. Then

$$P_\theta\{(X_{(c)}, X_{(n-c+1)}) \ni \theta\} = P_\theta(c \leq S(\theta) \leq n - c).$$

Since $S(\theta)\big|_\theta \stackrel{D}{=} S(0)\big|_{H_0}$, we have

$$P_\theta(c \leq S(\theta) \leq n - c) = P_{H_0}(c \leq S(0) \leq n - c).$$

Since the test is of level α, the RHS probability in the above is at least $1 - \alpha$. Thus the coverage probability of the random interval $[X_{(c)}, X_{(n-c+1)})$ is at least $1 - \alpha$. Hence $[X_{(c)}, X_{(n-c+1)})$ is a confidence interval for θ with confidence probability at least $1 - \alpha$.

3 Large sample test

Under the null hypothesis, $S \sim Binomial(n, \frac{1}{2})$. Now by DeMoiver-Laplace limit theorem $S^* = \frac{S - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$ is asymptotically $N(0, 1)$. Therefore, different tests can be performed in large samples using S^*. Consider testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$. Then the corresponding large sample test is non-randomized and rejects the null hypothesis if the observed value of S^* exceeds τ_α. Similarly, the large sample tests for the other hypotheses can also be constructed.

4 Sign test for quantiles

Suppose $p \neq 0.5$, that is $\theta(F)$ is the quantile of order p. Sign test can be still used to test $H_0 : \theta(F) = \theta_0$ against the usual alternatives. Properties like consistency and unbiasedness will be retained. However, the distribution of $S(\theta_0)$, in this case will be $Binomial(n, 1 - p)$ under the null hypothesis. Naturally the statistic for the large sample test will be $S^* = \frac{S - np}{\sqrt{np(1-p)}}$, which is asymptotically $N(0, 1)$ under the null hypothesis.

5 Paired Sample Sign Test

Suppose $(X_i, Y_i) i = 1, 2, .., n$ are sample observations from an unknown continuous bivariate distribution $F(x, y)$. Suppose the median of the distribution of difference $Z = X - Y$ is θ_Z. The objective is to test $H_0 : \theta_Z = \theta_0$ against the usual one sided and two sided alternatives. Naturally, $\theta_Z = \theta_0$ indicates that X observations tend to be θ_0 units larger than the corresponding Y observations. Thus statements about θ_Z give information about the relative locations of the marginal distributions. Then the appropriate procedure is simply Sign test based on the new sets of observations $Z_i = X_i - Y_i, i = 1, 2, .., n$.

6 Sample size determination

To perform a Sign test, we need a random sample. Suppose the objective is to determine a shift in the median. Consider testing $H_0 : \theta = 0$ against $H_a : \theta = \theta_1 (> 0)$ for specified θ_1. Now the test can be based on a small sample size or a large sample size. But often the observations are subject to some cost and time constraint. Therefore, the experimenter needs to choose the sample size, which will be sufficient to reach a decision. The usual technique is to determine n in such a way that the test has size α and power at the alternative $1 - \beta$, where α and β are specified in advance.
Consider the non-randomized level α Sign test for $H_a : \theta = \theta_1$, which rejects the null hypothesis if $S \neq k$, where k is such that $P_{H_0}(S \geq k) \leq \alpha$. Since S has a binomial distribution with parameters n and success probability 0.5, the above condition reduces to

$$\sum_{i=k}^{n} \binom{n}{i} (.5)^n \leq \alpha.$$

Suppose $p = P(X_1 > 0|\theta_1)$. Then the power of the above test becomes $\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^i$. Thus we need to determine sample size such that the conditions

$$\sum_{i=k}^{n} \binom{n}{i} (.5)^n \leq \alpha$$

and

$$\sum_{i=k}^{n}\binom{n}{i}p^i(1-p)^i \geq 1-\beta$$

are satisfied.

6.1 Calculation for various F

Suppose α is set at 6% and power at 80%. Then we get $n = 15$ and $k = 11$. Now we take $\theta_1 = .8$ and consider F as Normal, Cauchy and Logistic. For normal distribution, we get power .801 but for the Cauchy distribution it is .57 and for Logistic distribution it is .47. Thus for normal population, the required sample size is 15. However, we need more samples to achieve 80% power for the other distributions. Then for normal distribution, we get power .851 but for the Cauchy distribution it is .585 and for Logistic distribution it is .48. A similar exercise with $\alpha = .01$ gives $n = 25$ and $k = 18$. Then for normal distribution , we get power .851 but for the Cauchy distribution it is .585 and for Logistic distribution it is .48, which supports the need of further observations.

6.2 Approximate sample size

The determination of exact sample size is not easier in practice. We can use the large sample approximations to get a simple formula for sample size. Consider testing $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1 (> 0)$ for specified θ_1. The large sample test rejects the null hypothesis if $\frac{S}{n} > c$. Then the size and power requirements are

$$P_{H_0}(\frac{S}{n} > c) = \alpha,$$

and

$$P_{H_1}(\frac{S}{n} > c) = 1-\beta.$$

If σ_i^2 is the asymptotic variance of $\sqrt{n}\frac{S}{n}$ under H_i, i=0,1, then the above requirements can be equivalently expressed as

$$P_{H_0}(\sqrt{n}\frac{\frac{S}{n}-.5}{\sigma_0} > \frac{\sqrt{n}(c-.5)}{\sigma_0}) = \alpha,$$

and

$$P_{H_1}(\sqrt{n}\frac{\frac{S}{n}-p}{\sigma_1} > \sqrt{n}\frac{c-p}{\sigma_1}) = 1-\beta.$$

Then we get the approximate relations $c - .5 = \frac{\tau_\alpha \sigma_0}{\sqrt{n}}$, and $c - p = -\frac{\tau_\beta \sigma_1}{\sqrt{n}}$ and hence

$$n \approx \frac{(\sigma_1 \tau_\beta + \sigma_0 \tau_\alpha)^2}{(p-.5)^2},$$

where $\sigma_0^2 = \frac{1}{4}$ and $\sigma_1^2 = p(1-p)$. The same formula is valid for testing $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1 (< 0)$ for specified θ_1. For two sided alternative, we need to replace τ_α by $\tau_{\frac{\alpha}{2}}$.

6.3 Comparison of approximate sample size

For the purpose of comparison, we consider three distributions, namely, Normal, Cauchy and Logistic. The median for each distribution is taken as θ and scale parameter unity. Then we consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. For various choices of $\theta (> 0)$, we have computed the sample size by the derived formula with $\alpha = .05$ and $\beta = .2$. The nature of the approximate sample size is plotted in the next page.

6.4 Observations

The plot depicts the same fact as is observed in the exact numerical study. Normal distribution takes the lowest number of samples to reach the desired power level among the candidates. Logistic distribution takes the highest number of observations to reach 80% power. In addition, as we increase θ, the required sample size decreases for each candidate. This is a consequence of increasing power functions.

7 Application of Sign Test

7.1 Test of Trend: Cox-Stuart test

For data coming from some measurement processes, it is often desirable to check the presence of trend. That is, we are interested in knowing whether the observations depend on time.

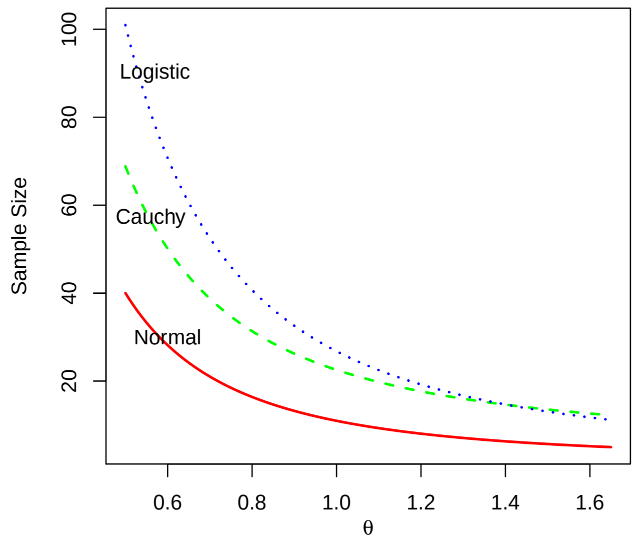


Figure 1: Approximate Sample size for different F

Suppose $X_i, i = 1, 2, .., n$ are iid observations from a continuous population F. The hypotheses for such tests are expressed as H_0 : Absence of trend in data against H_1 : Presence of upward trend in data.

Assume that the observations are sequentially observed and $X_1, X_2, .., X_n$ are ordered as they are observed. Also assume that n is even and define $c = \frac{n}{2}$(if n is odd, the middlemost observation, that is the $\frac{n+1}{2}$ th observation is removed). Then, the whole set of observations are grouped into $X_i, i = 1, 2, .., c$ and $X_i, i = c + 1, c + 2, .., n$. We club the observations into pairs $(X_i, X_{i+c}), i = 1, 2, .., c$. If an upward trend is present, the event $X_{i+c} > X_i$ is more probable than the event $X_{i+c} < X_i$ for every i. If $p = P(X_{i+c} > X_i)$, then the testing problem can be expressed as testing $H_0 : p = \frac{1}{2}$ against $H_1 : p > \frac{1}{2}$. Thus the problem can be looked up on as that for the Sign test.

Then the test statistic T is the number of pairs $(X_i, X_{i+c}), i = 1, 2, .., c$ for which $X_i < X_{i+c}$. That is $T = \sum_{i=1}^{c} I(X_i < X_{i+c})$. Thus the test statistic is nothing but the Sign test statistic

based on $X_{i+c} - X_i, i = 1, 2, .., c$. Naturally higher values of T indicates presence of an upward trend. Under the null hypothesis T has a binomial distribution with parameters c and success probability $\frac{1}{2}$. Then as usual depending on the given level α, the test can be constructed.

7.2 Test of correlation

Cox-Stuart test as discussed above can also be used to test for possible correlation. Suppose patients are given two drugs, one after another. Since the drugs are applied on the same patient, responses are correlated. Assume that a higher response indicates a favourable condition. Also assume that the paired response has a continuous bivariate distribution. Then the hypotheses in such a case can be expressed as H_0 : Absence of positive correlation against H_1 : Presence of positive correlation.

For details, assume that the response of the i the patient for the first drug(second drug) is $X_i(Y_i), i = 1, 2, .., n$. Order the pairs according to the increasing values of the X observations. For example if three pairs of observations are (5,3),(3,2) and (7,8), then the ordered pairs, ordered according to the magnitude of the first elements, are (3,2),(5,3) and (7,8). Then testing existence of positive (negative) correlation is equivalent to testing the presence of an upward(downward)trend in the ordered Y observations. Thus the test becomes the same as Cox-Stuart test of trend determination on the ordered Y observations.

8 Why use a Sign test?

The full information contained in the observations are not used in Sign test. Thus Sign test is less powerful, so one must use t test. But t test is based on normality. If the underlying distribution is non-normal, optimal test is rare to exist. In addition, the assumption about the underlying distribution is often instrumental. Thus Sign test is a safe option when there is any doubt about the normality of the underlying population though a sacrifice in the power.

Nonparametric Inference: Module 10[1]

What we provide in this module

- Wilcoxon Signed rank test
- Properties
- Exact null distribution
- Form of critical region
- Modification in the presence of ties

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Wilcoxon Signed Rank Test

This is another nonparametric alternative to Student's t test with the additional assumption of symmetry. This test is developed by Frank Wilcoxon(1945) but popularized by Sidney Siegel(1956). The procedure utilizes the signed rank of the observations and provides a distribution free test for location.

1.1 Assumptions & the hypothesis

Suppose $X_1, X_2, .., X_n$ are iid observations from a symmetric location family of distributions F. Then $F(x) = F(x - \theta(F))$, where $\theta = \theta(F)$ is a location parameter. F is assumed to be continuous and symmetric , that is, $F(x) + F(-x) = 1$ for all x. Under symmetry θ is, therefore, the median of F. Then our objective is to provide a test for $H_0 : \theta(F) = \theta_0$ against the usual one or two sided alternatives. However, without any loss of generality, we can take $\theta_0 = 0$

1.2 Signed Rank

The continuity assumption ensures that $P(X_i = 0) = 0$ for all i and that the observations are distinct with probability one. Suppose the observations are ranked in order of absolute value and ranked accordingly. Suppose R_i^+ is the rank of $|X_i|$ among $\{|X_1|, |X_2|, .., |X_n|\}$. Then signed rank of an observation is the rank of its absolute value multiplied by the sign of the original observation. If $Z_i = I(X_i > 0)$, then the signed rank of the i th observation is $Z_i R_i^+, i = 1, 2, .., n$. The signed-rank sum T is defined as the sum of the signed ranks, i.e. $T = \sum_{i=1}^n Z_i R_i^+$. We shall explain using the following example. Consider the following data with $n = 10$.

X_i	7	9	13	14	16	23	-11	2	-6	4
$\|X_i\|$	7	9	13	14	16	23	11	2	6	4
$Z_i R_i^+$	4	5	7	8	9	10	-6	1	-3	2

Then the signed rank sum is $T = 37$.

1.3 T as a test statistic

Note that X_i is expected to be larger under $\theta > 0$ than under $\theta = 0$. Thus a large(small) value of T implies that most of the large deviations from 0 are positive(negative). Therefore, a large(small) T is an indicator of positive(negative) θ. Then it seems reasonable to reject the null hypothesis against $H_a : \theta > 0 (H_a : \theta < 0)$ if T tend to be too large(or too small). Similarly too large and too small values of T indicates possible rejection of the null hypothesis against $H_a : \theta \neq 0$.

2 T is distribution free!

Now we shall show that the distribution of T does not depend on any F under the null hypothesis. Before we proceed further, we introduce the concept of antirank or inverse rank. If $\mathbf{R} = (R_1, R_2, .., R_n)$ is a rank vector, the antirank vector is $\mathbf{D}$, provided $\mathbf{R}\ o\ \mathbf{D} = (R_{D_1}, R_{D_2}, .., R_{D_n}) = (1, 2, .., n)$.

Consider an example with $n = 5$. Suppose $\mathbf{R} = (3, 2, 4, 1, 5)$ then

$$\mathbf{R}\ o\ \mathbf{D} = (R_{D_1}, R_{D_2}, .., R_{D_n}) = (1, 2, 3, 4, 5)$$

only when $\mathbf{D} = (4, 2, 1, 3, 5)$. Thus $\mathbf{D}$ is the antirank of $\mathbf{R}$.

Since T is the sum of few numbers from the set of first n integers, we can arrange the ranks in such a way that the i th rank comes in the i th position. If $\mathbf{D}$ is the antirank vector corresponding to $\mathbf{R}^+$, we can express T as $T = \sum_{i=1}^{n} iZ_{D_i}$.

Assume that $X_i \sim F(x)$, where $F(x) + F(-x) = 1$ for all x. Naturally F has median at the origin. Then for every x, it is easy to observe that

$$\begin{aligned} P(Z_i = 1, |X_i| \leq x) &= \frac{1}{2}(2F(x) - 1) \\ &= P(Z_i = 1)P(|X_i| \leq x) \end{aligned}$$

Since Z_i and $|X_j|$ are independent for any $i \neq j$, the desired independence follows. Now, being a function of $|X_i|, i = 1, 2, .., n$, $\mathbf{R}^+$ is independent of $Z_i, i = 1, 2, .., n$. Thus $Z_i, i = 1, 2, .., n$ and $D_i, i = 1, 2, .., n$ are independently distributed.

Here $\mathbf{D}$ is a random permutation over $\mathcal{P}$, the set of all n! permutations of $\{1, 2, .., n\}$. Again $Z_i \sim Bernoulli(\frac{1}{2})$ for every i. Thus

$$
\begin{aligned}
P(Z_{D_i} = z_i \ \forall i) &= \sum_{\mathbf{d} \in \mathcal{P}} P(Z_{d_i} = z_i \ \forall i | \mathbf{D} = \mathbf{d}) P(\mathbf{D} = \mathbf{d}) \\
&= \sum_{\mathbf{d} \in \mathcal{P}} P(Z_{d_i} = z_i \ \forall i) P(\mathbf{D} = \mathbf{d}) \\
&= \sum_{\mathbf{d} \in \mathcal{P}} (\frac{1}{2})^n P(\mathbf{D} = \mathbf{d}) \\
&= (\frac{1}{2})^n
\end{aligned}
$$

Thus $Z_{D_i}, i = 1, 2, .., n$ are iid random variables with Bernoulli($\frac{1}{2}$) distribution.

Hence T is the weighted sum of iid random variables $Z_{D_i}, i = 1, 2, ..n$. Under the symmetry about origin, $Z_{D_i}, i = 1, 2, ..n$ are iid Bernoulli($\frac{1}{2}$) variables. Thus under $H_0 : \theta = 0$, distribution of $Z_{D_i}, i = 1, 2, ..n$ and hence distribution of T is independent of any F. Thus T is exactly distribution free under the null hypothesis. Therefore tests based on T are exactly nonparametric.

3 Exact distribution of T

Consider the representation $T = \sum_{i=1}^n iZ_i$, where Z_i are iid Bernoulli($\frac{1}{2}$). Then T takes the values from 0 to $\frac{n(n+1)}{2}$ and under the symmetry of the underlying population

$$P(T = t) = \frac{b(t, n)}{2^n} \ t = 0, 1, 2, .., \frac{n(n+1)}{2},$$

where $b(t, n) = \#(z_1, z_2, ..., z_n) : \sum_{i=1}^n iz_i = t$ and z_i are either 0 or 1. Although no closed form expression exists for the distribution of T, we can compute the distribution for smaller values of n using the above definition.

We can calculate b(t, n) using the recursion equation $b(t,n) = b(t,n-1) + b(t-n,n-1)$, where with $b(t,n) = 0$ for $t < 0$ or $t > \frac{n(n+1)}{2}$ and $b(0,n) = b(1,n) = 1$ for all n. The recursion relation follows from the fact that

$$\begin{aligned} b(t,n) &= \{\#(z_1, z_2, \cdots, z_{n-1}) : \sum_{i=1}^{n-1} iz_i = t-n\} \\ &+ \{\#(z_1, z_2, \cdots, z_{n-1}) : \sum_{i=1}^{n-1} iz_i = t\} \end{aligned}$$

We give the calculation of the distribution for small values of n.

Suppose n=2, then possible values of T are 0,1,2 and 3. Then $b(0,2) = 1 = b(1,2)$. Now using the recursion relation $b(2,2) = b(2,1) + b(0,1) = 0 + 1 = 1$. Again applying the recursion, we get $b(3,2) = b(3,1)+b(1,1) = 0+1 = 1$. Then we have the following distribution:

t	$P(T = t)$
0	$\frac{1}{4}$
1	$\frac{1}{4}$
2	$\frac{1}{4}$
3	$\frac{1}{4}$

Next assume n=3, then possible values of T are 0,1,2,..,6. Then $b(0,3) = 1 = b(1,3) = 1$. Now using the recursion relation and the values of $b(.,2)$, we get $b(2,3) = b(2,2)+b(-1,2) = 1+0 = 1$, $b(3,3) = b(3,2) + b(0,2) = 1+1 = 2$, $b(4,3) = b(4,2) + b(1,2) = 0 + b(1,2) = 1$, $b(5,3) = b(5,2) + b(3,2) = 0 + 1$ and $b(6,3) = b(6,2) + b(3,2) = 0 + 1 = 1$. Then we have the following distribution:

t	0	1	2	3	4	5	6
$P(T = t)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

3.1 Symmetry of the distribution of T

If we look at the distribution for $n = 3$, we can observe symmetry about $n(n+1)/4 = 3$ Define $\mu_0 = n(n+1)/4$, then

$$\begin{aligned} P(T = t + \mu_0) &= P(\sum_{i=1}^{n} iZ_i = t + \mu_0) \text{ (as } Z_i \overset{iid}{\sim} Bernoulli(\tfrac{1}{2})) \\ &= P(\sum_{i=1}^{n} iZ_i = \mu_0 - t) \text{ (as } Z_i \overset{D}{=} 1 - Z_i \text{ for all } i) \end{aligned}$$

This holds for every $t \in \{0, 1, 2, .., n(n+1)/2\}$. Thus the distribution of T is symmetric about $\mu_0 = n(n+1)/4$.

3.2 Summary measures of the distribution of T

Since the distribution of T is symmetric about $\mu_0 = n(n+1)/4$, we get $E(T) = n(n+1)/4$. To find the variance, we use the representation $T = \sum_{i=1}^{n} iZ_i$, where Z_i are iid Bernoulli($\frac{1}{2}$) variables. Then

$$\begin{aligned} Var(T) &= Var(\sum_{i=1}^{n} iZ_i) \\ &= \sum_{i=1}^{n} i^2 Var(Z_i) \end{aligned}$$

Since $Var(Z_i) = \frac{1}{4}$ and $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$, we get $Var(T) = \frac{n(n+1)(2n+1)}{24}$.

4 Different Tests

T has a discrete distribution and hence tests based on it will be randomized. We list below the tests for different alternatives. For the alternative $H_a : \theta > 0$, a size α test can be expressed as

$$\phi^0 = I(T > T_\alpha) + aI(T = T_\alpha),$$

where T_α is such that $E_{H_0}\phi^0 = \alpha$.
For the alternative $H_a : \theta < 0$, a size α test can be expressed as

$$\phi^0 = I(T < T_{1-\alpha}) + aI(T = T_{1-\alpha}),$$

where $T_{1-\alpha}$ is such that $E_{H_0}\phi^0 = \alpha$.

However, for the alternative $H_a : \theta \neq \theta_0$, we note that the distribution of T is symmetric about $\mu_0 = \frac{n(n+1)}{4}$ under the null hypothesis. Hence a size α test can be expressed as

$$\phi^0 = I(|T - \mu_0| > T_{\frac{\alpha}{2}}) + aI(|T - \mu_0| = T_{\frac{\alpha}{2}})$$

with $E_{H_0}\phi^0 = \alpha$.

4.1 p values

Suppose T_{obs} is the observed value of T. Then for the alternative $H_a : \theta > 0$, one can report the one sided p value is $P_{H_0}(T \geq T_{obs})$. We accept the null hypothesis if this p value exceeds α.For the alternative $H_a : \theta < 0$, corresponding one sided p value is $P_{H_0}(T \leq T_{obs})$ and we accept the null hypothesis if this p value exceeds α. But for the alternative $H_a : \theta \neq 0$, the two sided p value is $2min\{P_{H_0}(T \geq T_{obs}), P_{H_0}(T \leq T_{obs})\}$. Thus we reject the null hypothesis if this p value does not exceed α.

4.2 Presence of ties

Presence of zero and tied observations undermine the validity of the test. Pratt(1959) recommended a modification of T in such a situation. He suggested to use the zeros and average ranks for the tied observations for the modification of the usual statistic.
To be specific, suppose $0 < u_1 < u_2 < ... < u_m$ be the distinct absolute magnitudes of the data $X_1, X_2, .., X_n$. Suppose f_0 is the frequency of zeroes in the data. $f_i^+(f_i^-)$ is the frequency of positive(negative) u_i and $f_i = f_i^+ + f_i^-$ is the total frequency of frequency of u_i in the data. Then Pratt suggested to use the statistic $T^* = \sum_{i=1}^m w_i f_i^+$, where

$w_j = f_0 + f_1 + .. + f_{j-1} + (f_j + 1)/2$. However, only large sample tests are suggested based on the asymptotic normality of the standardised statistic.

Nonparametric Inference: Module 11[1]

What we provide in this module

- Representation of T in terms of positive Walsh averages
- Consistency of Signed rank test
- Unbiasedness of Signed rank test
- Confidence interval using the cut offs
- Asymptotic normality & large sample tests
- Test for symmetry
- Sample size determination
- Comparing Signed rank test and Sign test

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Representation in terms of Walsh averages

Suppose $X_1, X_2, .., X_n$ are iid observations from a symmetric location family of distributions F. Then $F(x) = F(x - \theta(F))$, where $\theta = \theta(F)$ is a location parameter. F is assumed to be continuous and symmetric , that is, $F(x) + F(-x) = 1$ for all x. Under symmetry θ is, therefore, the median of F. Define Walsh averages $W_{ij} = \frac{X_i + X_j}{2}, 1 \leq i \leq j \leq n$. Then T can be expressed as the number of positive Walsh averages, i.e. $T = \sum_{j=1}^{n} \sum_{i=1}^{j} I(\frac{X_i+X_j}{2} > 0)$.

Observe that for $j > i$, $X_{(i)} < X_{(j)}$ and hence $I(X_{(i)} + X_{(j)} > 0) = 1$ if $X_{(j)} > 0$. If $X_{(i)} < 0$, then $X_{(i)} + X_{(j)} > 0$ implies that $|X_{(i)}| < X_{(j)}$ for $X_{(j)} > 0$. Thus $I(X_{(i)} + X_{(j)} > 0) = 1$ iff $|X_{(i)}| < X_{(j)}$ and $X_{(j)} > 0$. For i=j, $I(X_{(i)} + X_{(j)} > 0) = I(X_{(j)} > 0)$ and hence $I(X_{(j)} > 0) = 1$ if $X_{(j)} > 0$.

Hence, for $X_{(j)} > 0$, in $\sum_{i=1}^{j} I(X_{(i)} + X_{(j)} > 0)$, only those i will contribute for which $|X_{(i)}| < X_{(j)}$. Thus $\sum_{i=1}^{j} I(X_{(i)} + X_{(j)} > 0)$ is the signed rank of $X_{(j)}$ for $X_{(j)} > 0$. Then $R_j^+ = \sum_{j=1}^{i} I(X_{(i)} + X_{(j)} > 0)$ for $X_{(j)} > 0$. Equivalently we can write the above as $I(X_{(j)} > 0)R_j^+ = \sum_{i=1}^{j} I(X_{(i)} + X_{(j)} > 0)$. Summing the both sides over $j = 1$ to n, the desired expression follows as $\sum_{j=1}^{n} I(X_{(j)} > 0)R_j^+$ is nothing but $\sum_{j=1}^{n} Z_j R_j^+$.

1.1 Representation in terms of U statistic

Considering the representation in terms of positive Walsh averages, we have

$$\begin{aligned}
T &= \#(X_i, X_j) : \frac{X_i + X_j}{2} > 0, i \leq j \\
&= \#(X_i) : X_i > 0 + \#(X_i, X_j) : X_i + X_j > 0, i < j \\
&= \binom{n}{1} U_1 + \binom{n}{2} U_2,
\end{aligned}$$

where U_1 is a U statistic with kernel $I(X_1 > 0)$ and U_2 is a U statistic with kernel $I(X_1 + X_2 > 0)$.

2 Consistency of Signed rank Test

Consider testing $H_0 : \theta = 0$ against $H_a : \theta > 0$ using T. Signed rank test can be equivalently performed in terms of $\frac{T}{\frac{n(n+1)}{2}}$. From the convergence of U statistic, we find for any θ,

$$U_1 \xrightarrow{P} P(X_1 > 0)$$

and

$$U_2 \xrightarrow{P} P(X_1 + X_2 > 0).$$

Thus

$$\frac{T}{\frac{n(n+1)}{2}} \xrightarrow{P} P(X_1 + X_2 > 0).$$

Now define $\mu(F) = P(X_1 + X_2 > 0)$. Then , under $\theta = 0$,

$$\begin{aligned} \mu(F) &= EI(X_1 + X_2 > 0) \\ &= EE\{I(X_1 + X_2 > 0)|X_2\} \\ &= E(1 - F(-X_2)) = \frac{1}{2} \end{aligned}$$

Since for any i, $X_i \sim F(x - \theta) \Leftrightarrow X_i - \theta \sim F(x)$, the event $X_1 + X_2 > 0$ under $\theta = 0$ is a subset of that under $\theta > 0$ and hence $P_\theta(X_1 + X_2 > 0) > \frac{1}{2}$ for $\theta > 0$. Thus $\mu(F) >$ or $= \frac{1}{2}$ as $\theta > 0$ or $\theta = 0$ and hence Signed rank test is consistent against the alternative $H_a : \theta > 0$. Consistency against the other alternatives can also be proved.

3 Sign Test is unbiased

Consider testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$ and the representation $T(\theta) = \#(X_i, X_j) : \frac{X_i + X_j}{2} > \theta, i \leq j$. Assume $\theta_0 = 0$ then the power function can be expressed as

$$\begin{aligned} E_\theta \phi^0 &= P_\theta(T(0) > T_\alpha) + aP_\theta(T(0) = T_\alpha) \\ &= aP_\theta(T(0) > T_\alpha - 1) + (1 - a)P_\theta(T(0) > T_\alpha) \end{aligned}$$

Since for any i, $X_i \sim F(x-\theta) \Leftrightarrow X_i - \theta \sim F(x)$, $T(0)$ is expected to be larger under any θ than under $\theta = 0$. Thus under any $\theta > 0$, $P_\theta(T(0) > k) \leq P_{\theta=0}(T(0) > k)$ for any $k > 0$. Thus $E_\theta \phi^0 \geq P_{\theta=0}(T(0) > T_\alpha) + aP_{\theta=0}(T(0) = T_\alpha)$ Since the distribution of $T(\theta)$ under any θ is the same as that of $T(0)$ under $\theta = 0$, we find that the RHS of above equals $E_{\theta_0}\phi^0 = \alpha$. Thus we find that $E_\theta\phi^0 \geq \alpha$ for any $\theta \geq \theta_0$ and hence unbiasedness follows. Similarly unbiasedness for the other tests can be proved.

4 Confidence interval for θ

We can use the cut off of Signed rank test to get a confidence interval of θ with confidence coefficient at least $(1-\alpha)$. Now $T(0)$ can be looked upon as a sum of $m = \frac{n(n+1)}{2}$ positive Walsh averages. Then the acceptance region of the non randomized level α test for $H_0 : \theta = 0$ against $H_a : \theta \neq 0$ can be expressed as $c \leq T(0) \leq m - c$, where c is such that $P_{H_0}(c \leq T(0) \leq m - c) \geq 1 - \alpha$. Then from the theory of order statistics, we find that

$$T(\theta) \leq m - c \Leftrightarrow \theta \geq W_{(c)}$$

and

$$T(\theta) \geq c \Leftrightarrow \theta < W_{(m-c+1)},$$

where $W_{(i)}$ are the ordered Walsh averages. Thus

$$c \leq T(\theta) \leq m - c \Leftrightarrow W_{(c)} \leq \theta < W_{(m-c+1)}.$$

Hence $P_\theta\{(W_{(c)}, W_{(m-c+1)}) \ni \theta\} = P_\theta(c \leq T(\theta) \leq m - c)$ Since $T(\theta)\big|_\theta \stackrel{D}{=} T(0)\big|_{\theta=0}$, we have

$$P_\theta(c \leq T(\theta) \leq m - c) = P_{H_0}(c \leq T(0) \leq m - c) \geq 1 - \alpha.$$

Hence $[W_{(c)}, W_{(m-c+1)})$ is a confidence interval for θ with confidence probability at least $1-\alpha$.

4.1 Finding Walsh averages

Thus confidence interval depends on the ordered Walsh averages. Now we shall discuss how to obtain such averages in practice. Suppose $n = 4$ and the observations $X_i, i = 1, .., 4$ are

-2,5,8,13. Then we have $m = 10$ Walsh averages $\frac{X_i+X_j}{2}, i \leq j$ as below

	X_1	X_2	X_3	X_4
X_1	-2			
X_2	1.5	5		
X_3	3	6.5	8	
X_4	4.5	9	10.5	13

Thus we get $\mathbf{W} = (-2, 1.5, 5, 3, 6.5, 8, 4.5, 9, 10.5, 13)$ and hence the ordered values can be obtained.

5 Asymptotic normality

Under the null hypothesis, $T = \sum_{i=1}^{n} iZ_i$ is the sum of independent but non identical random variables. Then under the null hypothesis, as $n \to \infty$,

$$T^* = \frac{T - E(T)}{\sqrt{Var(T)}} \xrightarrow{D} N(0,1)$$

provided Liapounov's condition is satisfied. Liapounov's condition will be satisfied if

$$R_L = \frac{\sum_{i=1}^{n} E|iZ_i|^3}{\{\sum_{i=1}^{n} E|iZ_i|^2\}^{3/2}} \to 0$$

as $n \to \infty$. Since $E(Z_i^2) = E(Z_i^3) = \frac{1}{2}$, it is straightforward to show that the above condition is satisfied.

5.1 Large sample test

Since asymptotic normality holds, different tests can be performed in large samples using T^*. Consider testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$. Then the corresponding large sample test is non-randomized and rejects the null hypothesis if the observed value of T^* exceeds τ_α. Similarly, the large sample tests for the other hypotheses can also be constructed.

6 Application: Test for symmetry

Symmetry of the underlying distribution can also be tested using Wilcoxon signed-rank statistics. Suppose $X_i, i = 1, 2, .., n$ are iid observations from an unknown distribution F. Assume that F is a continuous distribution. To test H_0 F is symmetric with known median θ_0. That is under the null hypothesis, $X_i - \theta_0 \overset{D}{=} \theta_0 - X_i$ for each i. We suggest to use T and the related critical regions for the situation.

If H_0 is accepted, it is concluded that F is symmetric with median θ_0. However, rejection of H_0 is difficult to explain. Consider testing H_0 against all alternatives and suppose H_0 is rejected. Then either F is symmetric with median $\neq \theta_0$ or F is asymmetric with median θ_0 or F is asymmetric with median $\neq \theta_0$. However, such a broad conclusion , in general, lacks conclusiveness.

7 Paired Sample Wilcoxon Signed rank Test

Suppose $(X_i, Y_i) i = 1, 2, .., n$ are sample observations from an unknown continuous bivariate distribution $F(x, y)$. If the median of the distribution of difference $D = X - Y$ is θ_D, the objective is to test $H_0 : \theta_D = 0$ against the usual one sided and two sided alternatives. Naturally, $\theta_D = \theta_1 (> 0)$ indicates that X observations tend to be θ_1 units larger than the corresponding Y observations. Thus statements about θ_D give information about the relative locations of the marginal distributions. Then the appropriate procedure is simply the Wilcoxon Signed rank test based on the new sets of observations $D_i = X_i - Y_i, i = 1, 2, .., n$.

8 Signed Rank test & Sign test

Signed rank test is a nonparametric competitor of Student's t test. So which test is best? Of course the test with higher power is preferable. Now we shall give a comparative picture of the power of two tests. Consider n=10 observations from a $N(\theta, 1)$ population. To test $H_0 : \theta = 0$ against $H_a : \theta > 0$ at size $\alpha = .056$. Then the power function of Signed Rank test

is $P_\theta(T > T_\alpha) + aP_\theta(T = T_\alpha)$. If$\pi = P_\theta(X_1 > 0) = \Phi(\theta)$, then the power function of Sign test is $P_\theta(S > S_\alpha) + aP_\theta(S = S_\alpha)$, where $S \sim Binomial(10, \pi)$. For $n = 10$ and $\alpha = .056$ we find $T_\alpha = 43$ and $a = .25$. Again for the Sign test we find $S_\alpha = 7$ and $a = .0112$. Then we draw the two power curves on the same graph paper for varying θ. But the power for the Signed Rank test is obtained through simulation.

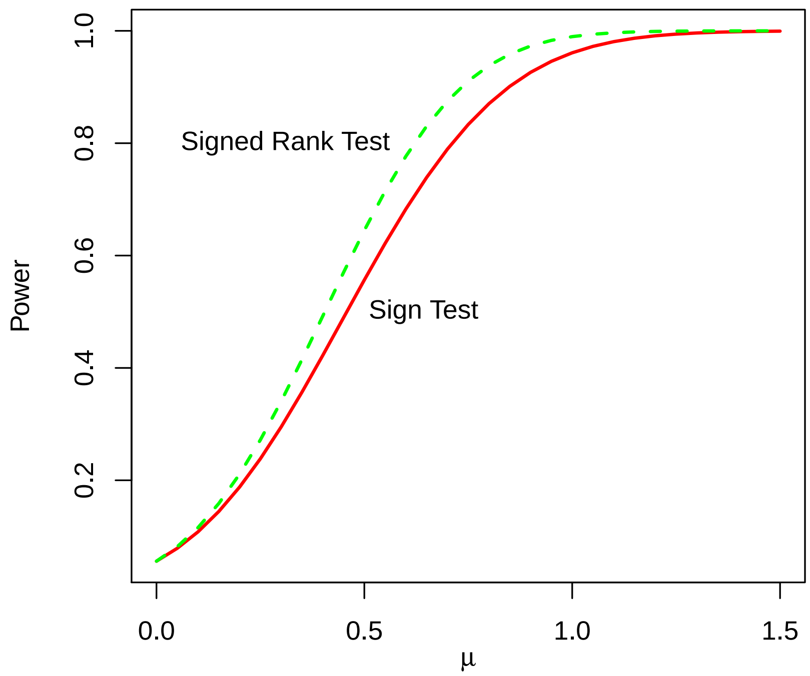


Figure 1: Nature of power for varying θ(indicated by μ)

8.1 Observation & Recommendation

The power curve for Sign test is always below than that of the Signed rank test. The reason is that, unlike Sign test, Signed rank test incorporates symmetry of the underlying distribution

through the test statistic. Thus Signed rank test is more powerful. Hence, if the underlying distribution is continuous and symmetric(can be checked through the histogram), Signed rank test is the most appropriate option. It is always a safe option when there is any doubt about the normality of the underlying symmetric population.

Nonparametric Inference: Module 12[1]

What we provide in this module

- Goodness-of-fit problems
- Pearsonian chi square tests
- Kolmogorov-Smirnov tests
- Properties : Consistency & Distribution free
- Some other goodness-of-fit tests

[1] Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 What is goodness of fit?

Often the underlying distribution of the data is not known. Traditional assumption of normality makes the situation weak in such cases and therefore the experimenter wish to know whether the distribution is of a given form. Thus the objective is to know which distribution "fits" good to the observed data. Tests for determining the underlying distribution, which is a good fit to the observed data are known as goodness-of-fit tests.

1.1 A motivating example

For better understanding, let us consider an example, where a random sample of size 9 is taken from a continuous distribution on [0,1] as

$$.01,.11,\ .23,\ .34,\ .55,\ .65\ ,\ .77\ ,.78,.89$$

Naturally, the sample size , that is, 9 is too small to assume normality. Again the observations are taken from a distribution with bounded support. Thus normality assumption will only be a forceful assumption. Suppose from a previous idea, it is known that the distribution can be assumed to be a $Beta(.5, .5)$ distribution. Then the usual methods can not be applied to test the hypothesis that the observations are coming from a $Beta(.5, .5)$ distribution.

1.2 Another example

Suppose the number of misprints per page of a book of 350 pages are represented in the form of the following grouped distribution:

No. of misprints/page	0	1	2	3	≥ 4
Number of pages	250	84	16	5	0

Here the number of observations is quite large to assume normality. But the variable is discrete with small number of categories. It is well known that the number of misprints/page has, in general, a Poisson distribution. But the mean, i.e. the only parameter, is not specified in advance. Then with the usual methods, testing the hypothesis that the observations are coming from a Poisson distribution with unknown parameter is not possible.

1.3 A further example

Now consider a data on the heights of 122 students in a certain class in the form of the grouped distribution given below:

Height(cm.)	154-159	159-164	164-169	169-174	174-179
# students	4	32	54	27	5

For the above data, the number of observations is quite large to assume normality. But the data is presented in the form of a frequency distribution. Moreover, the distribution of height is known to be normal. But even under the assumption of normality, the parameters are not given. Thus we need to develop further methods to check that the observations are coming from a normal distribution.

1.4 Related observations

From these examples, it is observed that:

- We often have data from an unknown continuous distribution with support other than the whole real line.
- We can have data from a known discrete distribution but with unspecified parameter(s).
- We can also have data from a known continuous distribution with unspecified parameter(s).

In these examples the data is given in either in the form of a frequency distribution or in the full and the objective was to test whether a given distribution fits the data well.

2 The hypothesis for goodness of fit

Suppose $X_i, i = 1, 2, .., n$ are iid observations from a distribution with DF $F(x)$. Then in the goodness-of-fit problems the objective is to test the null hypothesis

$$H_0 : F(x) = F_0(x) \ \forall x$$

against

$$H_a : F(x) \neq F_0(x) \text{ for at least one x.}$$

Now, we have two possibilities: $F_0(x)$ is completely known or $F_0(x)$ is not completely specified, i.e. some parameters remain unknown. For completely known $F_0(x)$, the null hypothesis is simple but in the other case the null hypothesis is composite. However, in each case the alternative is composite. Now we develop a number of tests for such hypotheses.

3 Pearsonian Chi-square test

Suppose, we have data in the form of a frequency distribution. For discrete distributions, the categories are natural(i.e. how many observations are zero or 1 or 2 etc?). But for continuous data, the experimenter prepares his/her own frequency distribution. Obviously conversion of data in the form of a grouped frequency distribution incurs a loss in information. Naturally, the lower the number of categories, the higher is the loss. Then, it is reasonable to classify the data set in to a higher number of categories.

Suppose n observations are available from the distribution of X. Let the range of X be divided into mutually exclusive and exhaustive sets $A_i, i = 1, 2, .., k$ with f_i as the number of observations classified in A_i. Then $\sum_{i=1}^{k} f_i = n$. Now define $p_i = P_{H_0}(X \in A_i), i = 1, 2, .., k$ as the probability that an X observation falls in A_i under the null hypothesis.

3.1 $F_0(x)$ is completely specified

For completely specified $F_0(x)$, p_i are completely known under the null hypothesis with $\sum_{i=1}^k p_i = 1$. Now the joint distribution of $(X_1, X_2, .., X_{k-1})$ is k-1 variate multi-nomial with parameters $(n, p_1, p_2, .., p_{k-1})$. Since $E(f_i) = np_i, i = 1, 2, .., k-1$, under the null hypothesis f_i are expected to be closer to np_i for each i. Then departure from the null hypothesis can be measured by $f_i - np_i$. Pearson's Chi-square test uses the statistic $U_k = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$ to measure the overall discrepancy.
It can be shown that under the null hypothesis, $U_k \xrightarrow{D} \chi^2_{k-1}$ as $n \to \infty$. Thus under the null hypothesis, the asymptotic distribution is independent of any F. Therefore, test provided by U_k is nonparametric. Now the higher is the discrepancy , the larger is the value of U_k. Therefore, a higher value of the statistic indicates departure from the null hypothesis. Thus, the large sample test rejects the null hypothesis if $U_k > \chi^2_{k-1,\alpha}$ at $100(1-\alpha)\%$ level of significance.

3.2 $F_0(x)$ is not completely specified

Suppose F_0 is specified except some parameters. Let the unknown parameters of F_0 be $(\theta_1, \theta_2, .., \theta_r)$. Then $p_i = P_{H_0}(X \in A_i), i = 1, 2, .., k$ are specified except for some parameters. In fact, p_i depends on the unknown $(\theta_1, \theta_2, .., \theta_r)$. Thus, in such a case $p_i = p_i(\theta_1, \theta_2, .., \theta_r)$, a function of the unknown parameters. Then we need to estimate p_i from the data to measure discrepancies. Suppose $\hat{\theta}_i, i = 1, 2, ..r$ be the ML estimates based on the data. Then $\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, .., \hat{\theta}_r), i = 1, 2, .., k$ are the ML estimates of $p_i, i = 1, 2, .., k$. Pearson's Chi-square statistic, in this situation is taken as $V_k = \sum_{i=1}^k \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}$. As earlier, a higher value of V_k indicates possible departure from the null hypothesis.
Now , it can be shown, as earlier that under the null hypothesis, $V_k \xrightarrow{D} \chi^2_{k-r-1}$ as $n \to \infty$. Thus under the null hypothesis, the asymptotic distribution is independent of any F. Therefore, test provided by V_k is asymptotically nonparametric. Since a higher value of V_k indicates departure from the null hypothesis, a right tailed test based on V_k is appropriate.

Thus, the large sample test rejects the null hypothesis if $V_k > \chi^2_{k-r-1,\alpha}$ at $100(1-\alpha)\%$ level of significance.

3.3 Chi-square test: Some facts

Now we shall discuss some important facts related to the above test. First of all, note that the choice of k is subjective, too small k fails to capture the features of the underlying distribution. Thus the experimenter needs to classify data in to as many categories as possible to gather more information about the underlying distribution. Often the data is available in the form of a grouped distribution. If np_i(or its estimate) is less than 5 for some i, the corresponding class is pooled with one or more neighbour classes so that the expected frequency for the combined class is at least 5. However, in such a case, degrees of freedom becomes number of classes after combining minus 1 minus the number of parameters estimated.

4 Kolmogorov-Smirnov tests

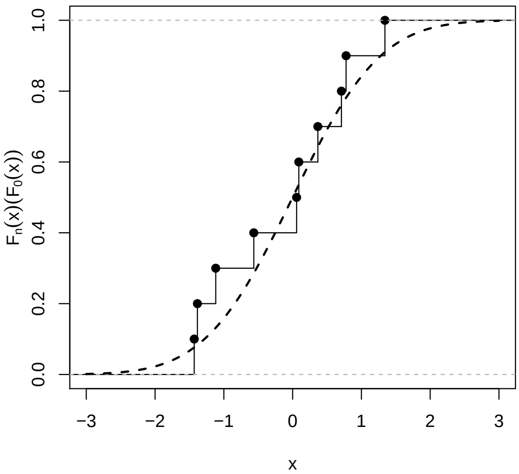
Next we discuss another test, which needs full data i.e. data is not available in the form of a grouped frequency distribution. Assume that F_0 is continuous and completely specified. Since, we need to measure the discrepancy between $F(x)$, the actual distribution and $F_0(x)$, a postulated known distribution, we use an estimate of $F(x)$. From the theory of U statistic, the empirical DF $F_n(x)$ is the U statistic for the estimation of $F(x)$. Then it is known that $E(F_n(x)) = F(x)$ and for each fixed x,

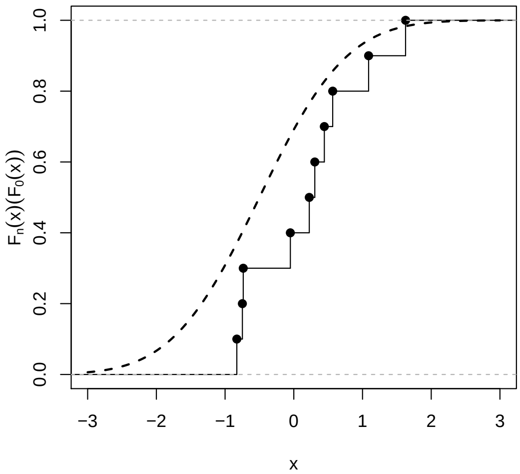
$$\sqrt{n}\{F_n(x) - F(x)\} \xrightarrow{D} N\{0, F(x)(1-F(x)\}.$$

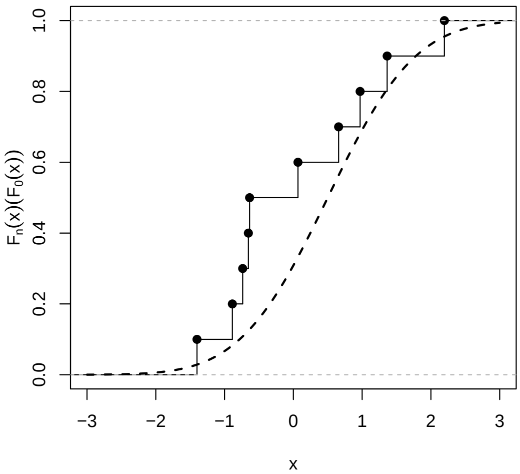
as $n \to \infty$.

4.1 Kolmogorov-Smirnov distance metric

Now to develop a measure of discrepancy, we consider the plots of ecdf based on n iid observations from $F(x)$ and DF $F_0(x)$. Naturally ecdf is a step function whereas DF is continuous. For demonstration, we plot ecdf $F_n(x)$ based on 10 observations from $N(0,1)$ and take $F_0(x)$ as N(-.5,1),N(0,1) and N(.5,1) , respectively. All these are given in the following plots.







Consider the figures and look at the vertical distance between $F_n(x)$ and $F_0(x)$ for fixed x. Observe that the distance is a measure of departure from $F_0(x)$. However, for some x, the distance is lower and for some x, the distance is higher. Therefore, we consider the maximum of such distances to measure the distance from $F_0(x)$. Naturally higher values of maximum of such distances would indicate departure from $F_0(x)$. This suggests to use the distance metric $\sup_x(F_n(x) - F_0(x))$ or $\sup_x(F_0(x) - F_n(x))$, which are known as Kolmogorov-Smirnov(KS) statistics.

4.1.1 Kolmogorov-Smirnov test: Critical region

Suppose the alternative is $H_a : F(x) \geq F_0(x)$ for some x. Then our statistic is $D_n^+ = \sup_x(F_n(x) - F_0(x))$ and we reject H_0 if D_n^+ is too large. For the alternative is $H_a : F(x) \leq F_0(x)$ for some x we use $D_n^- = \sup_x(F_0(x) - F_n(x))$ and we reject H_0 if D_n^- is too large. However, for the alternative is $H_a : F(x) \neq F_0(x)$ for some x, we use $D_n = max(D_n^+, D_n^-) = \sup_x |F_n(x) - F_0(x)|$ and we reject H_0 if D_n is too large. The exact values of the cut off can be obtained from the tables of Owen(1962) for specific choices of n and level α.

4.1.2 Kolmogorov-Smirnov Large sample Test

Now it can be shown that $4nD_n^{+2} \xrightarrow{D} \chi_2^2$ as $n \to \infty$. Thus a large sample test for $H_0 : F(x) = F_0(x) \forall x$ against $H_a : F(x) \geq F_0(x)$ for some x rejects the null hypothesis if $4nD_n^{+2} > \chi^2_{2,\alpha}$. Since $D_n^+ \stackrel{D}{=} D_n^-$ due to symmetry, a similar test against the alternative $H_a : F(x) \leq F_0(x)$ for some x can be obtained. Again, it can be shown that $\lim_{n\to\infty} P(\sqrt{n}D_n \leq z) = G(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}$. Thus a large sample test for $H_0 : F(x) = F_0(x)$ for all x against $H_a : F(x) \neq F_0(x)$ for some x rejects the null hypothesis if $\sqrt{n}D_n > d_\alpha$, where d_α is a root of $G(d_\alpha) = 1 - \alpha$.

4.1.3 Kolmogorov-Smirnov test is nonparametric

Now it is natural to inquire whether the test based on Kolmogorov-Smirnov statistic is distribution free. For this assume $F(x) = F_0(x)\forall x$. Note that ecdf can be represented as

$F_n(x) = \frac{i}{n}$ if $X_{(i)} \leq x < X_{(i+1)}$ for $i = 0, 1, 2, .., n$ with $X_{(0)} = 0$ and $X_{(n+1)} = \infty$. Then D_n^+ can be expressed as

$$\begin{aligned}
\sup_x (F_n(x) - F_0(x)) &= \max_{0 \leq i \leq n} \sup_{X_{(i)} \leq x < X_{(i+1)}} (F_n(x) - F_0(x)) \\
&= \max_{0 \leq i \leq n} (\frac{i}{n} - F_0(X_{(i)})) \\
&= \max_{0 \leq i \leq n} (\frac{i}{n} - U_{(i)}),
\end{aligned}$$

where $U_{(i)}, i = 0, 1, 2, .., n$ are the order statistics for a random sample of size n from a R(0,1) distribution. Thus D_n^+ is a function of $U_{(i)}, i = 0, 1, 2, .., n$, where $U_{(0)} = 0$. Since the joint distribution of $U_{(i)}, i = 1, 2, .., n$ is independent of any F under the null hypothesis, distribution of D_n^+ does not depend on any F. Thus the test given by D_n^+ is exactly nonparametric.Since $D_n^+ \stackrel{D}{=} D_n^-$, the test given by D_n^- is also nonparametric. Again, $D_n = max(D_n^+, D_n^-)$, and hence the corresponding test is also nonparametric.

4.1.4 Consistency of Kolmogorov-Smirnov test

Tests based on KS statistic are consistent. For a proof consider the alternative $H_a : F(x) \neq F_0(x)$ for some x. Then an asymptotically size α test rejects the null hypothesis if $4nD_n^{+2} > \chi^2_{2,\alpha}$ or equivalently if $D_n^+ > \sqrt{\frac{\chi^2_{2,\alpha}}{4n}}$. Now consider the Glivenko-Cantelli lemma,that is,

$$\sup_x (F_n(x) - F(x)) \xrightarrow{P} 0.$$

Then using Glivenko-Cantelli lemma, it can be shown that

$$D_n^+ \xrightarrow{P} \sup_x (F(x) - F_0(x)).$$

Then

$$\begin{aligned}
\sup_x (F(x) - F_0(x)) &= 0 \text{ if } F = F_0 \\
&> 0 \text{ if} F > F_0
\end{aligned}$$

Thus from the theory of consistency developed earlier, the one sided test is consistent. Similarly the consistency of the other tests can also be established.follow.

5 Few other goodness-of-fit statistic

In this module we have discussed only two distance metrics. But there are a number of such metrics which can be used as goodness of fit statistic. Some examples include

- Cramer-Von Mises(CvM) statistic defined by

$$C_n = \int (F_n(x) - F_0(x))^2 dF_0(x)$$

- Anderson-Darling (AD) statistic defined by

$$A_n = \int \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

For either statistic, we reject the null hypothesis if the observed value of the statistic is large. Tests provided by these statistics are also nonparametric. The proof follows from the representations:

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n} (U_{(i)} - \frac{2i-1}{2n})^2$$

and

$$nA_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)[logU_{(i)} + log(1 - U_{(n-i+1)})],$$

where $U_{(i)}, i = 1, 2, .., n$ are, as before, the order statistics for a random sample of size n from a R(0,1) distribution. Thus C_n and A_n are all distribution free and hence the tests provided by C_n and A_n are all nonparametric.

6 An application: Test for normality

Kolmogorov-Smirnov test can also be applied to test the normality of the given data in small samples. or normality, we take $F_0(x)$ as the DF of a standard normal variable. However, for the purpose of comparison, we consider the standardised version of the data set(say X^*). Then application of Kolmogorov-Smirnov test on X^* can reveal the underlying normality. Therefore, the procedure can be applied to identify non normal data even in small samples.

What we provide in this module

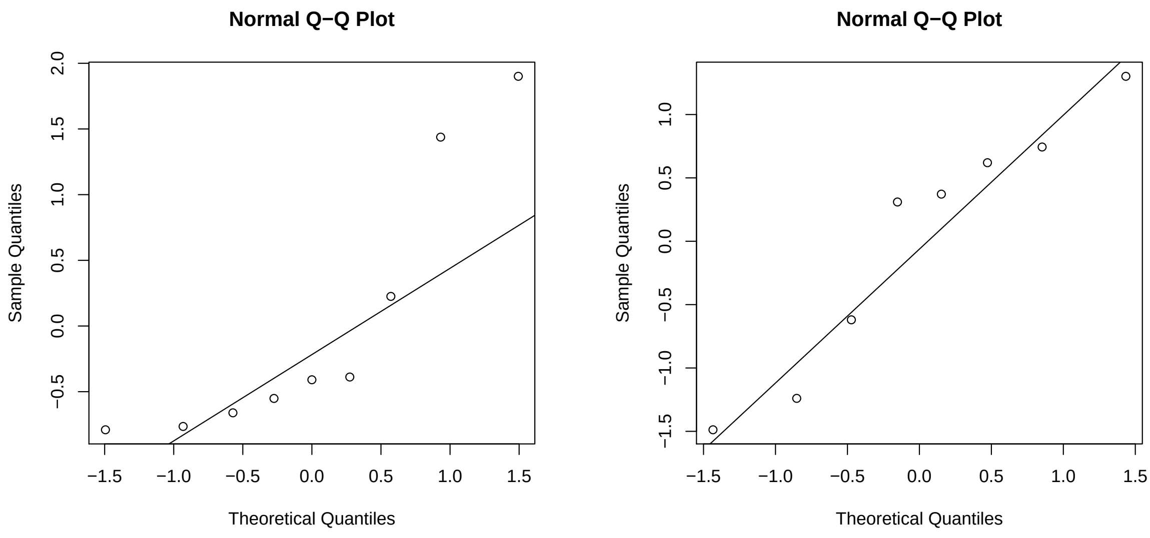
- Motivation for development
- Definition of Mann-Whitney U
- U as a measure of degree of separation of the two populations
- Null distribution of U
- Summary measures of U

1 Motivation

We start with an example with data from two independent populations. Consider the life-time(in 1000 hours)of bulbs corresponding to two different brands(say, Old and New)

Old	5.4	5.22	4.5	4.87	5.5	5.22	5.55	5.23	4.5	5.0
New	7.99	7.41	8.35	8.35	7.82	7.48	7.89	7.69		

Then it is natural to know which brand is performing better. The usual practice is to use Two sample t test using normality assumption. But we need normality for such a test and hence we perform some exploratory data analysis. We provide normal Q-Q plot and boxplot for sets of data.



(a) Old Brand (b) New Brand

Figure 1: Normal Q-Q plots of the standardised data

The QQ plot for Old brand data shows non normality whereas that for the New brand data reveals that the underlying distribution might be normal. Thus t test is not appropriate for this data. Again the box plot shows significant difference between the locations.

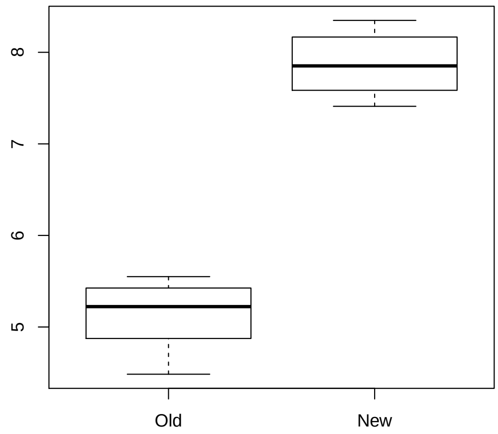


Figure 2: Box plot for the data

Since, deciding an appropriate distribution is subjective, applying a parametric test is not reasonable and hence, we need alternative procedures to judge the hypothesis.

2 The hypothesis

Suppose $X_i, i = 1, 2, .., n$ and $Y_j, j = 1, 2, .., m$ are independent observations from distributions F and G respectively. Then the null hypothesis can be expressed as

$$H_0 : F(x) = G(x) \text{ for all x .}$$

The most general two-sided alternative is $H_a : F(x) \neq G(x)$ for some x and the one-sided alternatives are $H_a : F(x) \geq G(x)$ for all x with strict inequality for some x or $H_a : F(x) \leq G(x)$ for all x with strict inequality for some x. Note that $F(x) \leq G(x)$ (or $F(x) \geq G(x)$) for all x with strict inequality for some x implies Y is either stochastically larger or smaller than X.

2.1 How to set the alternative?

Consider the data example. Then the objective is to know which brand is giving the more lifetime. That is for which brand the lifetime is expected to be higher. Suppose $X(Y)$ is the lifetime variable for old(new) brand bulbs. Then our interest is to know whether new brand bulbs are better , that is, $Y \overset{st}{>} X$. Thus for the given data the appropriate alternative should be $H_a : F(x) \geq G(x)$ for all x with strict inequality for some x. Depending on the need of the situation, the other alternatives are set.

2.2 Hypotheses under the location model

Suppose the alternative of interest is simply a difference in location(e.g. difference of average lifetime), then we assume $G(x) = F(x - \theta)\forall x$. That is the two populations differ only in location. Then ,

$$F(x) = G(x)\forall \text{ x} \Leftrightarrow \theta = 0$$

and

$$F(x) \geq\leq G(x)\forall \text{ x} \Leftrightarrow \theta \geq\leq 0.$$

Thus the testing problem reduces to $H_0 : \theta = 0$ against all alternatives. Note that under a location model

$$\theta \geq\leq 0 \Rightarrow \quad \text{the second population is shifted to the right(or left) of the first population.}$$

Now for a clear view of the location model, assume that $\frac{dF(x)}{dx} = f(x)$ exists for all x. Then the nature of $f(x - \theta)$ for different θ , that is, under different hypotheses can be graphically traced as below.

3 Mann-Whitney Statistic & properties

Suppose the X observations and Y observations are mixed together and ordered according to their magnitudes. Mann-Whitney statistic is based on the position of Y observations

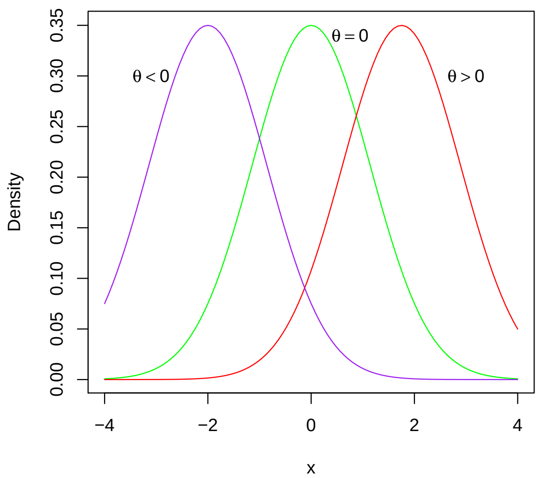


Figure 3: $f(x-\theta)$ for different θ

in the combined sample. If $\theta > 0(or < 0)$,then most of the Y observations are likely to be greater(or lower) than most of the X observations in the combined sample. Thus the number of times an X observation precedes a Y observation could be large(or small) in such a case. This suggests to use the statistic $U = \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i < Y_j)$ for our purpose. U was suggested by Mann and Whitney and are often called Mann-Whitney U .

3.1 U is distribution free

Let $(X_1, .., X_m, Y_1, .., Y_n)$ be the combined sample, where $N = m + n$. Now assume $F = G$, and consider the representation

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} I(F(X_i) < F(Y_j)).$$

Under $F = G$, $(F(X_1), .., F(X_m), F(Y_1), .., F(Y_n))$ can be treated as iid observations from a $R(0, 1)$ population. Then under $F = G$, U is a function of iid R(0,1) variables. Thus the distribution is independent of any F under $F = G$ and depends only on m and n.

Hence U becomes a distribution free statistic and therefore, tests based on U are exactly nonparametric.

4 U as a measure of degree of separation

First of all, we need to know the range of U. Consider two extreme cases, namely, all the X's are larger than Y's and all the Y's are larger than X's. In the former situation U is minimum ,i.e. 0 and in the latter case it is maximum , i.e. mn. For other configurations, $0 < U < mn$ and hence we get $0 \leq U \leq mn$. U actually measures the degree of separation of the populations. Naturally U takes the extreme values if the two populations are completely separated. With the well mixed observations, U takes the intermediate values. The following figures will be helpful to understand this. From Figure 4, it is easy to observe that for completely separated observations U is either highest or lowest and U takes the intermediate values depending on how different observations from the two populations are mixed.

4.1 Null Distribution of U

Define $p_{n,m}(u) = P_{H_0}(U = u),\ 0 \leq u \leq mn$. Now arrange the n X observations and m Y observations in increasing order of magnitude. Then under $F = G$, each of $\binom{m+n}{n}$ arrangements of n X's and m Y's are equally likely. If $N_{n,m}(u)$ is the number of arrangements such that the number of X's preceding Y's is u, then $p_{n,m}(u) = \frac{N_{n,m}(u)}{\binom{m+n}{n}}$. We shall derive a recursion relation for $N_{n,m}(u)$.

Now in the arranged sequence, the last value is either an X observation or an Y observation. If the largest value is an X observation, it does not alter the value of U. Then the number of arrangements with the remaining (n-1) X values and m Y values with an X value as the last element so that $U = u$ is $N_{n-1,m}(u)$. However, if the largest value is an Y observation, it is higher than n X's. Then the n X values and remaining m-1 Y values will give $U = u - n$. Then the number of arrangements with n X values and m-1 Y values with a Y value as the

last element so that $U = u - n$ is $N_{n,m-1}(u-n)$. Thus we get

$$N_{n,m}(u) = N_{n-1,m}(u) + N_{n,m-1}(u-n).$$

Since $p_{n,m}(u) = \frac{N_{n,m}(u)}{\binom{m+n}{n}}$, we get the following recursion relation:

$$p_{n,m}(u) = \frac{n}{m+n} p_{n-1,m}(u) + \frac{m}{m+n} p_{n,m-1}(u-n).$$

We note that $p_{n,m}(u) = 0$ if $u < 0$ and for $n \geq 1$, we have

$$\begin{aligned} p_{n,0} &= 1 \text{ if } \ u = 0 \\ &= 0 \text{ if } \ u > 0 \end{aligned}$$

Also for $m \geq 1$, we have

$$\begin{aligned} p_{0,m} &= 1 \text{ if } \ u = 0 \\ &= 0 \text{ if } \ u > 0 \end{aligned}$$

4.1.1 Finding Exact Distribution

For small m and n one can easily compute the distribution of U. We shall give few examples.

Suppose m=n=1, then $U = I(X_1 < Y_1) \sim Bernoulli(.5)$.

Next suppose n=2 and m=1, then $U = 0, 1, 2$. Thus

$$N_{2,1}(0) = N_{1,1}(0) + N_{2,0}(-2) = N_{0,1}(0) + N_{1,0}(-1) + 0 = 1.$$

$$N_{2,1}(1) = N_{2,1}(1) + N_{2,0}(-1) = N_{0,1}(1) + N_{1,0}(0) = 0 + 1 = 1.$$

Then $N_{2,1}(2) = 3 - 2 = 1$ and hence $P(U = u) = \frac{1}{3} \forall$ u.

Next suppose n=2 and m=2, then $U = 0, 1, .., 4$. Thus

$$N_{2,2}(0) = N_{1,2}(0) + N_{2,1}(-2) = N_{0,2}(0) + N_{1,1}(-1) = 1,$$

$$N_{2,2}(1) = N_{1,2}(1) + N_{2,1}(-1) = N_{0,2}(1) + N_{1,1}(0) = 0 + N_{0,1}(0) + N_{1,0}(-1) = 1,$$

and

$$N_{2,2}(2) = N_{1,2}(2) + N_{2,1}(0) = N_{0,2}(2) + N_{1,1}(1) + N_{1,1}(0) + N_{2,0}(-2) = 0 + N_{0,1}(1) + N_{1,0}(0) + 1 = 2.$$

In a similar way $N_{2,2}(3) = N_{2,2}(4) = 1$ and hence $P(U = u)$ can be obtained

4.2 U has a symmetric distribution

Note that under $F = G$, $X_i - Y_j \stackrel{D}{=} Y_j - X_i$ for every (i,j). Then

$$\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i < Y_j) \stackrel{D}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i > Y_j).$$

Thus under $F = G$, $U \stackrel{D}{=} mn - U$. Therefore, under $F = G$,

$$U - \frac{mn}{2} \stackrel{D}{=} \frac{mn}{2} - U$$

Hence the null distribution of U is symmetric about $\frac{mn}{2}$.

4.3 U in terms of placements

With n observations from F(x) and m observations from G(x), define $P_{(j)} = nF_n(Y_{(j)})$. $P_{(j)}$ is called the placement of $Y_{(j)}$ among the X observations. Then

$$\begin{aligned}
U &= \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_{(i)} < Y_{(j)}) \\
&= \sum_{j=1}^{m}\{\sum_{i=1}^{n} I(X_{(i)} < Y_{(j)})\} \\
&= \sum_{j=1}^{m}\{\#X's : X_{(i)} \leq< Y_{(j)}\} \\
&= n\sum_{j=1}^{m} F_n(Y_{(j)}) = \sum_{j=1}^{m} P_{(j)}
\end{aligned}$$

Thus U can be represented in terms of the placements.

4.4 Summary measures of U

Under the null hypothesis, U has a symmetric distribution and hence $E_{H_0}(U) = \frac{mn}{2}$. Note that $\frac{U}{mn}$ can be looked upon as a U statistic with the kernel $\theta = P(Y_1 > X_1)$. Then it is already proved that $\sigma_{10}^2 = E(1 - G(X_1))^2 - \theta^2$, $\sigma_{01}^2 = E\{F(Y_1)\}^2 - \theta^2$ and $\sigma_{11}^2 = \theta(1-\theta)$. Then under the null hypothesis $\theta = \frac{1}{2}$, $\sigma_{10}^2 = \frac{1}{12} = \sigma_{01}^2$ and $\sigma_{11}^2 = \frac{1}{4}$. Using these in the expression of exact variance of U statistic, we get that $Var(U|H_0) = \frac{mn(m+n+1)}{12}$.

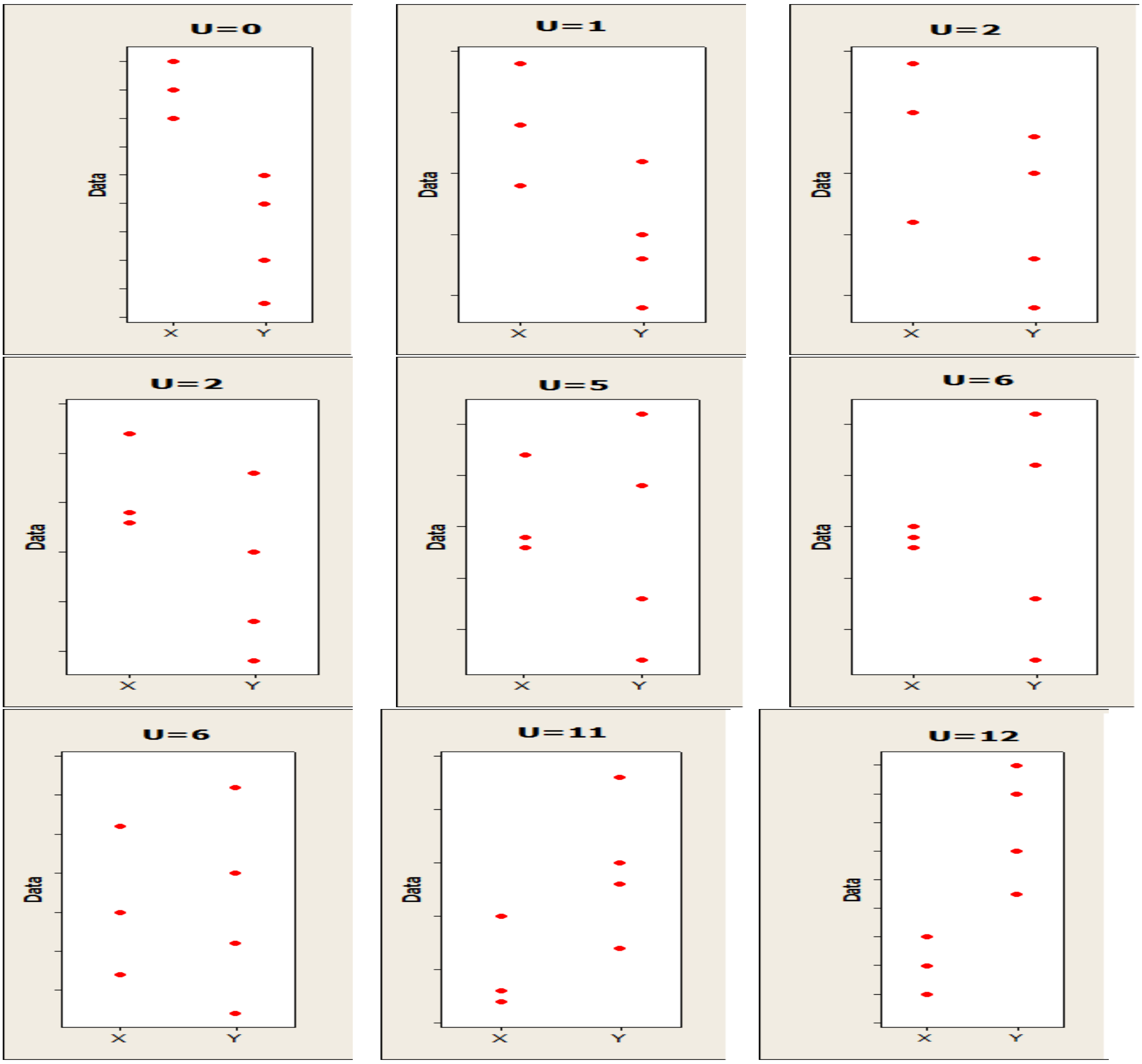


Figure 4: Degree of separation and the value of U

What we provide in this module

- Direct derivation of Variance of U
- Different tests: Small sample and large sample
- Modification for tied and ordinal data
- Consistency
- Confidence interval for median using Mann-Whitney cut offs

1 Variance of U: Direct approach

We have already discussed derivation of variance of U using the results of U statistics. But such a variance can be developed from the first principle. Define $D_{ij} = I(X_i < Y_j)$, then $U = \sum_{i=1}^{n} \sum_{j=1}^{m} D_{ij}$, then D_{ij} are not all independent and hence

$$\begin{aligned} Var(U) &= \sum_{ij} Var(D_{ij}) \\ &+ \sum_{i=1}^{n} \sum_{1 \le j \ne k \le m} Cov(D_{ij}, D_{ik}) \\ &+ \sum_{j=1}^{m} \sum_{1 \le i \ne k \le n} Cov(D_{ij}, D_{kj}) \\ &+ \sum_{1 \le i \ne g \le n} \sum_{1 \le j \ne k \le m} Cov(D_{ij}, D_{gk}) \end{aligned}$$

Evaluating each term, the final expression can be obtained. First of all note that $Var(D_{ij}) = p(1-p)$, where $p = P(X_1 < Y_1)$. Then

$$\begin{aligned} Cov(D_{ij}, D_{ik}) &= P(Y_j > X_i, Y_k > X_i) - p^2 \\ &= \int (1 - G(x))^2 dF(x) = p_1 - p^2, \end{aligned}$$

$$\begin{aligned} Cov(D_{ij}, D_{kj}) &= P(Y_j > X_i, Y_j > X_k) - p^2 \\ &= \int F^2(x) dG(x) = p_2 - p^2 \end{aligned}$$

and

$$\begin{aligned} Cov(D_{ij}, D_{gk}) &= P(Y_j > X_i, Y_k > X_g) - p^2 \\ &= p^2 - p^2 = 0. \end{aligned}$$

Now we have mn $Var(D_{ij})$, m(m-1) and n(n-1) covariance terms. Combining all these we get the final expression

$$Var(U) = mn\{p - (m + n - 1)p^2 + (m - 1)p_1 + (n - 1)p_2\}.$$

However, under $F = G$, $p = \frac{1}{2}$, $p_1 = p_2 = \frac{1}{3}$ and hence , we finally get $Var(U|H_0) = \frac{mn(m+n+1)}{12}$.

2 Different Tests

2.1 Small sample test

From the intuitive justification, we have the following tests. For the alternative $H_a : \theta > 0$, a size α test can be expressed as $\phi^0 = I(U > U_\alpha) + aI(U = U_\alpha)$, where U_α is such that $E_{H_0}\phi^0 = \alpha$. For $H_a : \theta < 0$, a size α test can be expressed as $\phi^0 = I(U < U_{1-\alpha}) + aI(U = U_{1-\alpha})$, where $U_{1-\alpha}$ is such that $E_{H_0}\phi^0 = \alpha$. However the distribution of S is symmetric about $\frac{mn}{2}$ under the null hypothesis and hence , a size α test for the two sided alternative $H_a : \theta \neq 0$ can be expressed as $\phi^0 = I(|U - \frac{mn}{2}| > U_{\frac{\alpha}{2}}) + aI(|U - \frac{mn}{2}| = U_{\frac{\alpha}{2}})$ with $E_{H_0}\phi^0 = \alpha$. One can also report the p values of the above tests to judge the strength of the observed data. If U_{obs} is the observed value of U, the concerned one sided p values are either $P_{H_0}(U \geq U_{obs})$ or $P_{H_0}(U \leq U_{obs})$. The two sided p value is $2min\{P_{H_0}(U \geq U_{obs}), P_{H_0}(U \leq U_{obs})\}$. In any case, the null hypothesis is rejected if this p value does not exceed α.

2.2 Large sample distribution & test

Using the results of U statistic, under the null hypothesis, $U^* = \frac{\frac{U}{mn} - \frac{1}{2}}{\sqrt{\frac{m+n+1}{12mn}}}$ is asymptotically $N(0,1)$. Asymptotic normality holds even for moderate (n,m) and for better understanding , we have plotted the mass function of U for various (n,m) and superimposed the corresponding normal density in the following figure.

Since, the asymptotic normality of U^* is sample sizes, different tests can be performed in large samples using U^*.Consider testing $H_0 : \theta = 0$ against $H_a : \theta > 0$.Then the corresponding large sample test is non-randomized and rejects the null hypothesis if the observed value of U^* exceeds τ_α. Similarly, the large sample tests for the other hypotheses can also

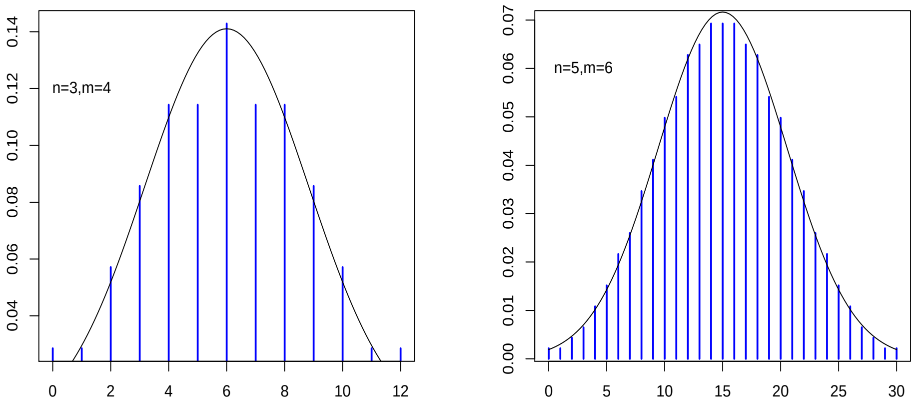


Figure 1: Normal Approximation for small (n,m)

be constructed.

2.3 Continuity Correction

The large sample distribution of standardised U gives an approximation of a discrete distribution by a continuous distribution. That is, we are approximating the area of a histogram by the area under a continuous curve. Thus, in general, $P_{H_0}(U \leq x)$ will be underestimated by the approximation $\Phi(\frac{x-mn/2}{\sqrt{mn(m+n+1)/12}})$. So we need a modification, commonly known as the continuity correction, which uses the fact that

$$P_{H_0}(U \leq x) \approx \Phi(\frac{x - mn/2 + 1/2}{\sqrt{mn(m+n+1)/12}}).$$

We explain the improvement by a relevant figure. For illustration, we have taken m=4,n=3 and plotted the corresponding DF of U(i.e. the step function). We have also plotted the normal DF without continuity correction(the continuous curve) and the normal DF with continuity correction(the dotted thick curve). Naturally, the continuity correction covers more region. Thus, better approximation is expected after correction for continuity.

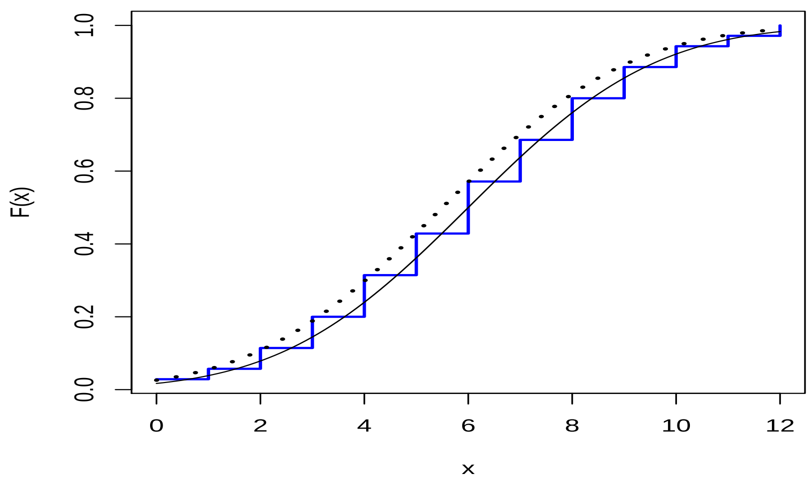


Figure 2: Continuity Correction with n=4,m=3

3 Modification for tied observations

We have already assumed continuity of F and G. But in practice ties can occur within either of the samples. However, in such a case we get a unique value of U. But if one or more X observation is tied with one or more Y observations, given definition of U is not applicable. In such a case we redefine D_{ij} as

$$\begin{aligned} D_{ij} &= +1 \text{ if } X_i < Y_j \\ &= 0 \text{ if } X_i = Y_j \\ &= \ \text{ if } X_i > Y_j \end{aligned}$$

Then the Mann-Whitney U reduces to $U_T = \sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}$. If we define $p^+ = P(D_{ij} = 1)$, $p^- = P(D_{ij} = -1)$ and $p_0 = P(D_{ij} = 0)$, then $E(D_{ij}) = p^+ - p^-$ so that $E(U_T) = mn(p^+ - p^-)$. The null hypothesis of interest in this case is $H_0 : p^+ = p^-$. Now, $Var(D_{ij}) = 1 - p_0 - (p^+ - p^-)^2$ and as earlier, $Var(U_T)$ can be obtained. However, the variance of U_T conditional upon the observed ties under H_0 can be calculated as $Var(U_T|\mathbf{t}) = \frac{mn(N+1)}{12}(1 - \frac{\sum_{k=1}^{s}(t_k^3 - t_k)}{N^3 - N})$,

where $N = m+n$, and t_i is the multiplicity of the i th tie. Naturally $t_i = 1 \forall i$ in the non tied situation. Then for moderate N, large sample normality of the statistic $\frac{U_T - 0}{\sqrt{Var_{H_0}(U_T|\mathbf{t})}}$ holds and hence large sample tests can be performed.

3.1 Mann-Whitney test for ordinal data

Adjustment to incorporate ties enables Mann-Whitney test applicable for ordinal data. We start with an example. Consider the following data on grades obtained by a group of patients assigned Treatment 1 and Treatment 2(A being the highest).

Trt 1	A	B	C
Trt 2	B	C	

The objective is naturally to compare the performance of two treatments. However, the responses are tied as there are 2 observations with B and 2 observations with grade C. Now we shall calculate $3X2$ D_{ij} values in the following matrix:

Y→ X↓	C	B	A
C	0	1	1
B	-1	0	1

The sum of the elements (i.e 2) is the value of U_T. Also the tied ranks are of length $t_B = 2$ and $t_C = 2$ Thus the modified statistic can be calculated for the purpose.

4 Consistency of Mann-Whitney Test

Consider testing $H_0 : \theta = 0$ against $H_a : \theta > 0$. Now the test can be equivalently expressed in terms of $T = \frac{U}{mn}$. Then from the convergence of U statistic

$$T \xrightarrow{P} P(X_1 < Y_1).$$

If we define $Y_1^0 = Y_1 - \theta$, then

$$Y_1 \sim F(x-\theta) \Leftrightarrow Y_1^0 \sim F(x).$$

Thus we get $P(X_1 < Y_1) = P(X_1 < Y_1^0 + \theta)$. Now, if we define $\mu(F) = P(X_1 < Y_1^0 + \theta)$ then $\mu(0) = \frac{1}{2}$ is the value of $\mu(F)$ under $F = G$. Hence

$$\begin{aligned} \mu(F) &= \frac{1}{2} \text{ if } \theta = 0 \\ &> \frac{1}{2} \text{ if } \theta > 0 \end{aligned}$$

Thus the Mann-Whitney Test is consistent against the alternative $H_a : \theta > 0$. Consistency against the other alternatives can also be proved.

5 Confidence interval for θ

We can use the cut off of Mann-Whitney test to get a confidence interval of θ with confidence coefficient at least $(1-\alpha)$. Consider the acceptance region of the non randomized level α Sign test for $H_0 : \theta = 0$ against $H_a : \theta \neq 0$. Note that under any θ, $X_i, i = 1, 2, .., n$ and $Y_j, j = 1, 2, .., m$ are iid observations from F. Now define

$$U(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\{I(Y_j - X_i > \theta).$$

Then Mann-Whitney test statistic is $U(0)$. Now, due to symmetry, the acceptance region can be expressed as $c \leq U(0) \leq mn - c$, where c is such that $P_{H_0}(c \leq U(0) \leq mn - c) \geq 1 - \alpha$. Let us arrange the mn differences $Y_j - X_i$ in increasing order and denote them by $D_{(1)} \leq D_{(2)} \leq ... \leq D_{(mn)}$. Then

$$U(\theta) \leq n - c \Leftrightarrow \theta > D_{(c)}.$$

Again

$$U(\theta) \geq c \Leftrightarrow \theta < D_{(mn-c+1)}.$$

Thus $c \leq U(\theta) \leq n-c$ is equivalent to $D_{(c)} \leq \theta < D_{(mn-c+1)}$. Then $P_\theta\{(D_{(c)}, D_{(mn-c+1)}) \ni \theta\} = P_\theta(c \leq D(\theta) \leq n-c)$. Since $U(\theta)\big|_\theta \stackrel{D}{=} U(0)\big|_{H_0}$, we have

$$P_\theta(c \leq U(\theta) \leq n-c) = P_{H_0}(c \leq U(0) \leq n-c).$$

Since the test is of level α, the RHS probability in the above is at least $1-\alpha$. Thus the coverage probability of the random interval $[D_{(c)}, D_{(mn-c+1)})$ is at least $1-\alpha$. Hence $[D_{(c)}, D_{(mn-c+1)})$ is a confidence interval for θ with confidence probability at least $1-\alpha$.

6 Determination of Sample size

Consider testing $H_0 : \theta = 0$ against $H_a : \theta = \theta_1(> 0)$ for specified θ_1. The experimenter should design the experiment in such a way that the difference θ_1 can be identified with high probability but with minimum the sample size. The usual technique is to determine (n, m) in such a way that the test has size α and power at the alternative $1-\beta$, where α and β are specified in advance. However determination of exact sample size is not easier in practice. We can use the large sample normal approximations to get a tractable formula for sample size.

Consider testing $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1(> 0)$ for specified θ_1. Then following Noether(1987), the approximate total sample size N can be expressed as

$$N \approx \frac{(\tau_\beta + \tau_\alpha)^2}{12\lambda(1-\lambda)(p-.5)^2},$$

where $p = P(Y_1 > X_1)$.

Naturally, p depends on unknown F. Now, for the purpose of comparison, we consider three distributions, namely, Normal, Cauchy and Laplace. For the first population, we consider N(0,1), Cauchy(0,1) and Laplace(0,1) distributions and for the second population, we take the same distributions but with location parameter θ. Then we consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. We have computed the sample size for various choices of $\theta(> 0)$, by the derived formula taking $\alpha = .05$ and $\beta = .2$. The nature of the approximate sample size is plotted in the following figure.

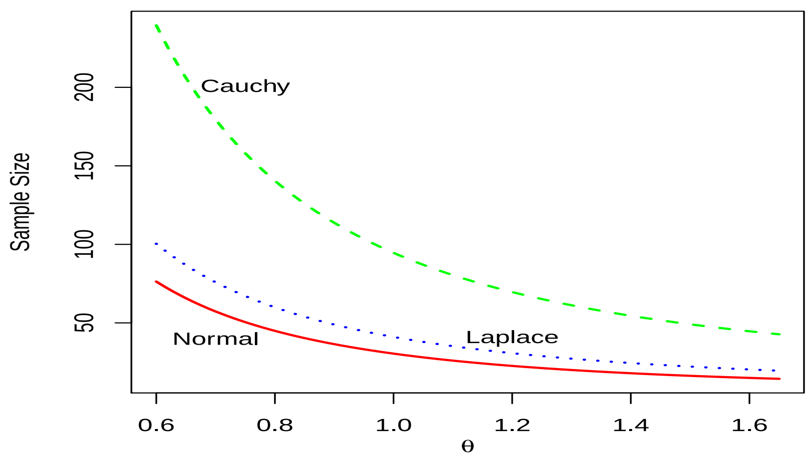


Figure 3: Approximate sample size for different F

Thus we observe that Normal distribution takes the lowest number of samples to reach the desired power level among the candidates and Cauchy distribution takes the highest number of observations to reach 80% power. In addition, as we increase θ, the required sample size decreases for each candidate.

Nonparametric Inference: Module 15[1]

What we provide in this module

- Wald-Wolfowitz run test with properties
- Different tests based on runs
- Kolmogorov-Smirnov two sample test based on ECDF
- Test for randomness

[1] Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Test based on arrangements: The idea

Suppose $X_i, i = 1, 2, .., n$ and $Y_j, j = 1, 2, .., m$ are independent observations from distributions F and G respectively and our hypothesis of interest is expressed as $H_0 : F(x) = G(x)$ for all x. The most general two-sided alternative is $H_a : F(x) \neq G(x)$ for some x. For the test of general alternative, we can use the arrangement pattern of the observations to derive a meaningful test. Under the null hypothesis the two samples can be regarded as a single sample of size $N = m + n$ from a continuous but unspecified population. Then we have $\binom{N}{n}$ equally likely configurations of m X's and n Y's. The arrangement pattern gives information about the closeness of the underlying populations. For example, if the observations are well mixed, then we could think of identical populations. However, if there is a pattern in mixing, a deviation from the null is natural. Many statistical tests are based on the pattern of mixing in the combined arrangement of N observations.

2 Wald-Wolfowitz Runs test

This test is based on the theory of runs. A run is defined as a sequence of similar symbols preceded and followed by a different symbol or by no symbol at all. For example consider 4 x's any 3 y's arranged in increasing order of magnitude in the following way:

$$x \quad y \quad y \quad x \quad x \quad y \quad x$$

Then we have a run of one x , then a run of 2 y's, then a run of 2 x's, a run of one y and lastly a run of a single x. The run length of the first run is unity, that of the second is 2 and so on. As another example consider 3 x's any 3 y's arranged in increasing order of magnitude in the following way:

$$x \quad y \quad x \quad y \quad x \quad y$$

Then we have 6 runs each of length unity.

Runs test is based on the total number of runs in the combined arrangement of n X's

and m Y's. If the two samples are from the identical populations, the observations are expected to be well mixed. Thus R is expected to be too large. On the other hand if the underlying populations are different, observations from different populations tend to be separated. Hence R is expected to be too small in such a case. Thus small values of R are indicative of the different populations.

However R is only a measure of deviation from $F(x) = G(x) \forall x$ and it does not differentiate between $Y \overset{st}{>} X$ or $X \overset{st}{>} Y$. For example consider the configuration:

$$x \quad x \quad x \quad x \quad y \quad y \quad y$$

Then R=2 but all the X observations are less than Y observations. This might indicate $Y \overset{st}{>} X$ Consider another configuration:

$$y \quad y \quad y \quad y \quad x \quad x \quad x$$

Then again R=2 but the configuration indicates the possibility of $Y \overset{st}{>} X$. Thus runs test is appropriate for the two-sided alternative is $H_a : F(x) \neq G(x)$ for some x.

2.1 Distribution of R

Assume that $F(x) = G(x) \forall x$. Suppose the X observations are indicated by x's and Y observations are indicated by y's. Under the null hypothesis we have $\binom{N}{n}$ equally likely configurations of m X's and n Y's. Suppose R_1 is the number of runs of x's and R_2 is the number of runs of y's. We shall first obtain the joint distribution of (R_1, R_2). Suppose (r_1, r_2) is the observed value of (R_1, R_2). Then note that $r_1 = 1, 2, .., n$ and $r_2 = 1, 2, .., m$ but r_1 and r_2 can not differ by more than unity. Because then two runs with the same symbols have to be adjacent, which contradicts the definition of runs. If $r_1 = r_2 - 1$, the arrangement starts with a run of y's and for $r_2 = r_1 - 1$, the arrangement starts with a run of x's. However, if

$r_1 = r_2$, then the arrangement starts with either a run of x's or a run of y's. Thus, we define

$$\begin{aligned} c(r_1, r_2) &= 0 \text{ if } |r_1 - r_2| \geq 2 \\ &= 1 \text{ if } |r_1 - r_2| = 1 \\ &= 2 \text{ if } |r_1 - r_2| = 0 \end{aligned}$$

Now under the null hypothesis, each of $\binom{m+n}{n}$ arrangements of m x's and n y' are equally likely. Now to get r_1 runs of x's we need to put $r_1 - 1$ bars in available n-1 positions. And for each such choice, if we put $r_2 - 1$ bars in available m-1 positions, we get r_2 runs of y's. Then the total number of ways of getting r_1 runs of x's and r_2 runs of y's is $c(r_1, r_2)\binom{n-1}{r_1-1}\binom{m-1}{r_2-1}$. Then we have the joint pmf

$$P(R_1 = r_1, R_2 = r_2) = \frac{c(r_1, r_2)\binom{n-1}{r_1-1}\binom{m-1}{r_2-1}}{\binom{m+n}{n}},$$

for $r_1 = 1, 2, .., n$ and $r_2 = 1, 2, .., m$.

Next we shall obtain the marginal distributions. By definition

$$\begin{aligned} P(R_1 = r_1) &= \sum_{r_2} P(R_1 = r_1, R_2 = r_2) \\ &= \frac{c(r_1, r_2)\binom{n-1}{r_1-1}\binom{m-1}{r_2-1}}{\binom{m+n}{n}} \end{aligned}$$

and a simple algebra gives

$$P(R_1 = r_1) = \frac{\binom{n-1}{r_1-1}\binom{m+1}{r_1}}{\binom{m+n}{n}}, r_1 = 1, 2, .., n.$$

In a similar way $P(R_2 = r_2) = \frac{\binom{m-1}{r_2-1}\binom{n+1}{r_2}}{\binom{m+n}{n}}, r_2 = 1, 2, .., m$.

Now to obtain the distribution of $R = R_1 + R_2$, the total number of runs, we note that

$$P(R_1 + R_2 = r) = \sum_{r_1=0}^{r} P(R_1 = r_1, R_2 = r - r_1).$$

Now for even r, the only possible value of r_1 is $r/2$. Thus

$$P(R_1 + R_2 = r) = 2\frac{\binom{n-1}{r/2-1}\binom{m-1}{r/2-1}}{\binom{m+n}{n}}, r = 2, 4, ..$$

However, for odd r, either $r_1 = (r-1)/2$ or $r_1 = (r+1)/2$. Then we get

$$P(R_1 + R_2 = r) = \frac{\binom{n-1}{\frac{r-1}{2}}\binom{m-1}{\frac{r-3}{2}}}{\binom{m+n}{n}} + \frac{\binom{n-1}{\frac{r-3}{2}}\binom{m-1}{\frac{r-1}{2}}}{\binom{m+n}{n}}, r = 1, 3, 5, ..$$

2.2 Moments of (R_1, R_2)

We start from the following lemma. **Lemma:** *For any a, b and k*

$$\sum_{i=0}^{min(a,b-k)} \binom{a}{i}\binom{b}{k+i} = \binom{a+b}{a+k}.$$

Proof: It is easy to observe that
$\binom{a}{i}$=coefficient of $\frac{1}{t^i}$ in $(1+\frac{1}{t})^a$ and $\binom{b}{k+i}$=coefficient of t^{k+i} in $(1+t)^b$. Thus we get

$$\sum_i \binom{a}{i}\binom{b}{k+i} = \text{ coefficient of } t^k \text{ in } (1+\frac{1}{t})^a(1+t)^b = \binom{a+b}{a+k}.$$

This completes the proof.

Now assume that $m > n$. Then with the notation $r^{(g)} = \frac{r!}{(r-g)!}$ $r \geq g$, we get the factorial moment

$$\begin{aligned} E{R_1}^{(g)} &= \sum_{r_1=1}^{n} (r_1)^{(g)} P(R_1 = r_1) \\ &= (m+1)^{(g)} \frac{\binom{m+n-g}{m}}{\binom{m+n}{n}} \end{aligned}$$

using the above lemma. In a similar way we get

$$E{R_2}^{(g)} = (n+1)^{(g)} \frac{\binom{m+n-g}{n}}{\binom{m+n}{n}}.$$

Now to obtain the expectation of R, we take $g = 1$ and get $E(R_1) = \frac{n(m+1)}{m+n}$. In a similar way, we get $E(R_2) = \frac{m(n+1)}{m+n}$. Hence we get $E(R) = E(R_1) + E(R_2) = 1 + \frac{2mn}{m+n}$.
Again taking $g = 2$, we get $ER_1(R_1 - 1) = \frac{n(n-1)}{(m+n)(m+n-1)}$ and hence we have

$$\begin{aligned} Var(R_1) &= ER_1(R_1 - 1) + E(R_1)(E(R_1 - 1)) \\ &= \frac{(m+1)^{(2)} n^{(2)}}{(m+n)(m+n)^{(2)}}. \end{aligned}$$

In a similar way we get $Var(R_2) = \frac{(n+1)^{(2)}m^{(2)}}{(m+n)(m+n)^{(2)}}$.

To obtain the variance of R, we need to compute the covariance between R_1 and R_2 and hence we start from the joint factorial moment, defined by

$$\mu^{(g,g)} = E(R_1 - 1)^{(g)}(R_2 - 1)^{(g)}.$$

Now by definition,

$$\mu^{(g,g)} = \sum_{r_1, r_2 \geq g+1} \frac{(r_1 - 1)!}{(r_1 - 1 - g)!} \frac{(r_2 - 1)!}{(r_2 - 1 - g)!} p(r_1, r_2),$$

where $p(r_1, r_2)$ is the joint pmf of (R_1, R_2). Then, again using the lemma we get

$$\mu^{(g,g)} = \frac{(n-1)^{(g)}(m-1)^{(g)}\binom{\bar{m}+\bar{n}}{\bar{n}}}{\binom{m+n}{n}},$$

where $\bar{r_1} = r_1 - g$, $\bar{r_2} = r_2 - g$, $\bar{n} = n - g$ and $\bar{m} = m - g$.

Now putting $g = 1$, we get $E(R_1-1)(R_2-1) = \frac{(n-1)(m-1)\binom{m+n-2}{\bar{n}}}{\binom{m+n}{n}}$ and hence using the relation, $Var(R) = Var(R_1) + Var(R_2) + 2cov(R_1, R_2)$, we get that $Var(R) = \frac{2mn(2mn-m-n)}{(m+n-1)(m+n)^2}$. For further details, we refer the reader to the book by Wilks(1962).

2.3 Different tests: small sample & large sample

From the intuitive justification, we reject the null hypothesis if $R < c$ for some c satisfying $P_{H_0}(R < c) < \alpha$. The test can also be performed in large samples. Now if we assume $\frac{n}{m+n} \to \lambda \in (0,1)$, then we have the following approximation $E(\frac{R}{m+n}) \approx 2\lambda(1-\lambda)$ and $Var(\frac{R}{\sqrt{m+n}}) \approx 4\lambda^2(1-\lambda)^2$. Then under the null hypothesis, if $min(m,n) \to \infty$ such that $\frac{n}{m+n} \to \lambda \in (0,1)$, then

$$R^* = \frac{R - 2N\lambda(1-\lambda)}{2\lambda(1-\lambda)\sqrt{N}} \xrightarrow{D} N(0,1),$$

where $N = m + n$. Since the large sample distribution of R^* is standard normal, a large sample test rejects the null hypothesis if $R^* < -\tau_\alpha$.

3 The Homogeneity Problem

We have already observed that Run test is insensitive to stochastic alternatives, like H_a : $F(x) \geq G(x)$ for all x with strict inequality for some x (i.e. $Y \overset{st}{>} X$) or $H_a : F(x) \leq G(x)$ for all x with strict inequality for some x (i.e. $X \overset{st}{>} Y$). If we can measure the discrepancy between $F(x)$ and $G(x)$, tests for one and two sided alternatives can be derived. From the theory of U statistic, the empirical DF's $F_n(x)$ and $G_m(x)$ are the U statistics for the estimation of DF's. If $F = G$, there should be a good agreement between $F_n(x)$ and $G_m(x)$ for all values of x. Some overall measure of discrepancy between these functions is then a natural test statistic. Then just as in Goodness-of-fit problems, we use the statistics $\sup_x(F_n(x) - G_m(x))$ or $\sup_x(G_m(x) - F_n(x))$. These statistics are nothing but the Kolmogorov-Smirnov(KS) distance function measuring the discrepancy between two continuous DF's.

4 KS two sample test

Kolmogorov-Smirnov(KS) two sample test of homogeneity is based on the KS distance function, defined above. Consider the alternative $H_a : F(x) \geq G(x)$ for all x. Then our statistic is $D^+_{n,m} = \sup_x(F_n(x) - G_m(x))$ and we reject H_0 if $D^+_{n,m}$ is too large. For the alternative is $H_a : F(x) \leq G(x)$ for all x, we use $D^-_{n,m} = \sup_x(G_m(x) - F_n(x))$ and we reject H_0 if $D^-_{n,m}$ is too large. However, for the alternative is $H_a : F(x) \neq G(x)$ for some x, we use $D_{n,m} = max(D^+_{n,m}, D^-_{n,m}) = \sup_x |F_n(x) - G_m(x)|$ and we reject H_0 if $D_{n,m}$ is too large.

4.1 KS test is nonparametric

Consider the statistic $D_{n,m} = \sup_x |F_n(x) - G_m(x)|$. Now assume $F = G$ and consider the new set of data, $F(X_i), i = 1, 2, .., n$ and $G(Y_j), j = 1, 2, .., m$. Note that $F_n(x)$ and $G_m(x)$ does not change for the new set of data. Now the transformed data are observations from a R(0,1) population and hence the distribution of $D_{n,m}$ under F=G depends only on m and n. Thus $D_{n,m}$ and hence $D^+_{n,m}$ and $D^-_{n,m}$ provides exactly nonparametric tests.

4.2 Consistency of KS test

Note that F_n and G_m are strongly consistent for F and G respectively. Thus $F_n - G_m$ is consistent for $F - G$ and hence $D_{n,m}$, $D^+_{n,m}$ and $D^-_{n,m}$ are all consistent estimators of the corresponding population quantities. Under F=G, all the population quantities are zero. On the other hand, it can be shown that $D_{n,m}$, $D^+_{n,m}$ and $D^-_{n,m}$ converge in probability to positive values. Thus KS test is consistent.

4.3 Large sample test

Now it can be shown that

$$4\frac{mn}{m+n}D^{+2}_{n,m} \xrightarrow{D} \chi^2_2$$

as $min(m,n) \to \infty$. Thus a large sample test for $H_0 : F(x) = G_(x)\forall x$ against $H_a : F(x) \geq G(x)$ for all x rejects the null hypothesis if

$$4\frac{mn}{m+n}D^{+2}_{n,m} > \chi^2_{2,\alpha}.$$

Again, it can be shown that

$$\lim_{min(n,m)\to\infty} P(\sqrt{\frac{mn}{m+n}}D_{n,m} \leq z) = H(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}.$$

Thus a large sample test for H_0 against $H_a : F(x) \neq G(x)$ for some x rejects the null hypothesis if $\sqrt{\frac{mn}{m+n}}D_{n,m} > d_\alpha$, where d_α is a root of $H(d_\alpha) = 1 - \alpha$.

5 Test for randomness

The most important application of runs test is to provide a distribution free test of randomness. We start with a real life example. Suppose the students entered a lecture theatre in a certain order. Then it is of interest to know whether students of the same gender come in together. If we denote male students by M and female students by F, then we get a sequence of symbols M and F. Naturally well mixed symbols tend to contradict that the students of

the same gender come in together or they enter the lecture theatre in a random fashion. Any pattern in the symbols possibly indicate lack of randomness. Thus our null hypothesis in this case is the randomness and the alternative is naturally lack of randomness. The arrangement of M and F gives a number of runs of these symbols. If the number of runs is too few or too large, some pattern is expected. Therefore, a contradiction to the random pattern is expected and hence we reject the null hypothesis in such a situation.

Formally if $X_i, i = 1, 2, .., n$ are observations from a continuous population F(x) then the null hypothesis can be expressed as $H_0 : X_1, X_2, .., X_n$ is a random sample. However, the alternative, that is , lack of randomness does not specify any particular pattern. Now to dichotomize the data(if the data is not naturally dichotomised) we use some statistic T based on all the observations. If we put the symbols L and U as the observation is lower or greater than T, we get an arrangement of two symbols. Naturally, we reject the null hypothesis, if the total number of runs is too small or too large. Using the large sample distribution of runs, a large sample test can also be performed.

Nonparametric Inference: Module 16[1]

What we provide in this module

- Concepts of concordance and discordance
- Kendall's τ with its estimate
- Properties
- Test of trend

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Measure of Association in parametric context

Suppose $(X_i, Y_i), i = 1, 2, .., n$ are iid observations from a continuous bivariate distribution $F(x, y)$ and the objective is to get some idea about the nature of dependence between X and Y. The most frequently used measure is the Pearson correlation coefficient, defined by

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

ρ is a scaled measure of degree of linear association and the direction of association. A large positive value(of course lees than unity) indicates that large values of X are paired with large values of Y and vice versa.

Although ρ is well accepted as an association measures but it suffers from a number of drawbacks. First of all, it is a measure of linear dependence between X and Y, i.e., if the actual relationship is actually or at least approximately linear, ρ is a good measure. But if the actual relationship is far from the linearity, ρ gives misleading results. For example, if $X \sim N(0, 1)$, then the covariance between X and $Y = X^2$ is zero, that is $\rho = 0$ and hence there is no linear relationship between X and Y. Then X and Y are uncorrelated, but there is actually an exact quadratic relationship. In addition, ρ heavily depends on the existence of the first few moments of the underlying distribution. Lastly, the value of ρ remains unchanged if the underlying variables are replaced by some linear functions with positive slope. That is, if the transformation is linear and order preserving then ρ does not change. However, if the transformation is non linear but order preserving, the value of ρ may be changed.

2 Concordance & Discordance

In order to get invariance under order preserved transformation, we must depend on relative magnitudes instead of absolute magnitudes and consequently, we introduce the concepts of concordance and discordance. Consider two pairs of observations (x, y) and (x', y'). These pairs are concordant(or in agreement) if they vary in the same direction, that is, if $x < x'$

whenever $y < y'$ or if $x > x'$ whenever $y > y'$.

Again the pairs (x, y) and (x', y') are said to be discordant (or in disagreement) if they vary in the opposite direction, that is, if $x > x'$ whenever $y < y'$ or if $x < x'$ whenever $y > y'$. In other words, if the line joining these points has a positive slope, the points are concordant and discordant otherwise. Naturally, concordance or discordance property is preserved under an order preserving transformation. The concept of concordant and discordant pairs are explained through a figure below.

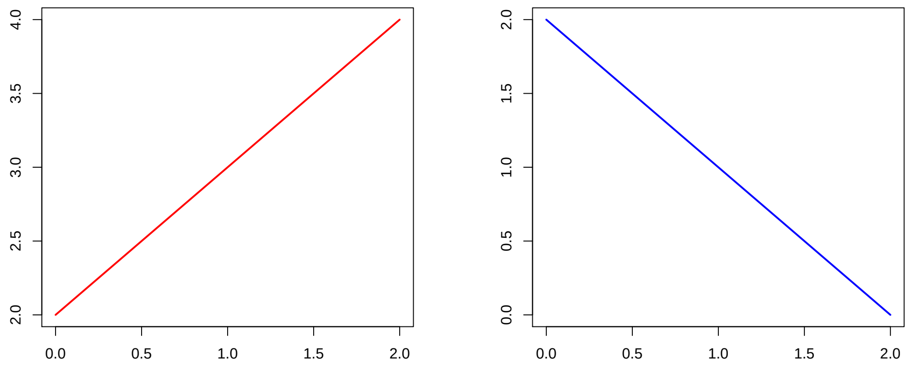


Figure 1: Concordant & Discordant pairs

2.1 Concordance and Discordance probabilities

For any two independent pairs (X_1, Y_1) and (X_2, Y_2), the probability of concordance is defined as

$$\begin{aligned}
\pi_c &= P\{[(X_1 > X_2)\&(Y_1 > Y_2)] \cup [(X_1 < X_2)\&(Y_1 < Y_2)]\} \\
&= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} \\
&= P\{\frac{(Y_2 - Y_1)}{(X_2 - X_1)} > 0\}
\end{aligned}$$

Thus π_c can be looked upon as the probability that the straight line joining the random points (X_1, Y_1) and (X_2, Y_2) has a positive slope. Similarly, the probability of discordance is defined as

$$\pi_d = P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$$

Thus the probabilities of concordance and discordance are invariant under order preserved transformations and hence a measure of association can be developed by the combinations of these probabilities.

2.2 Kendall's τ

Since perfect association is either perfect concordance or perfect discordance, a difference between π_c and π_d measures the extent of association. The difference $\pi_c - \pi_d$ is defined as the Kendall's τ coefficient. If the marginal distributions are assumed to be continuous then $P(X_1 - X_2 = 0) = 0 = P(Y_1 - Y_2 = 0)$ and hence we get $\pi_c + \pi_d = 1$. Thus we can express τ as $\tau = 2\pi_c - 1 = 1 - 2\pi_d$. Naturally, τ does not depend on the moments of the underlying distributions and hence is always defined. Since τ is the difference of two probabilities, we have $-1 \leq \tau \leq 1$.

Now we shall explain independence in terms of τ. Under independence, $X_1 - X_2$ and $Y_1 - Y_2$ are independent and hence

$$X_1 \stackrel{D}{=} X_2 \text{ and } Y_1 \stackrel{D}{=} Y_2.$$

Thus under independence

$$\begin{aligned} \pi_c &= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} \\ &= P\{(X_2 - X_1)(Y_1 - Y_2) > 0\} \\ &= P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} = \pi_d \end{aligned}$$

Hence $\tau = 0$ under independence.

However, the converse is not in general true and hence we get a similarity with ρ. For ex-

ample, consider a bivariate normal population with correlation coefficient ρ. Then

$$\text{Independence} \Leftrightarrow \tau = 0 \Leftrightarrow \rho = 0.$$

To be specific, suppose $(X_i, Y_i), i = 1, 2$ are iid observations from a $N_2(\theta_X, \theta_Y, \sigma_X^2, \sigma_Y^2, \rho)$ population. Then the joint distribution of $U = \frac{X_1 - X_2}{\sigma_X \sqrt{2}}$ and $V = \frac{Y_1 - Y_2}{\sigma_Y \sqrt{2}}$ is a standard bivariate normal distribution with correlation coefficient ρ. Then using the properties of bivariate normal distribution, we get

$$p_c = P(UV > 0) = \frac{1}{2} + \frac{1}{\pi} \sin^{-1} \rho.$$

Then $\tau = \frac{2}{\pi} \sin^{-1} \rho$ and hence we get the desired relationship with ρ.

2.3 Estimating Kendall's τ

Suppose $(X_i, Y_i), i = 1, 2, .., n$ are iid observations from a bivariate continuous distribution F. If we define $sgn(u)$ as

$$\begin{aligned} sgn(u) &= +1 \text{ if } u > 0 \\ &= 0 \text{ if } u = 0 \\ &= -1 \text{ if } u < 0, \end{aligned}$$

then we get

$$\begin{aligned} sgn(X_1 - X_2) sgn(Y_1 - Y_2) &= +1 \text{ if } (X_1 - X_2)(Y_1 - Y_2) > 0 \\ &= -1 \text{ if } (X_1 - X_2)(Y_1 - Y_2) < 0 \end{aligned}$$

and consequently

$$\begin{aligned} E\{sgn(X_1 - X_2) sgn(Y_1 - Y_2)\} &= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} \\ &- P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \\ &= \pi_c - \pi_d = \tau. \end{aligned}$$

Motivated by the above, we define the symmetric kernel

$$\phi((X_1, Y_1), (X_2, Y_2)) = sgn(X_1 - X_2)sgn(Y_1 - Y_2)$$

with degree 2. Then the corresponding U statistic is

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \{sgn(X_i - X_j)sgn(Y_i - Y_j)\} \\ &= \frac{1}{\binom{n}{2}} \{\#\text{concordant pairs} - \#\text{discordant pairs}\} \\ &= \frac{1}{\binom{n}{2}} (C - D), \end{aligned}$$

where $C(D)$ is the number of concordant(discordant) pairs. U is defined as Kendall's sample tau measure.

2.4 τ as correlation coefficient

Although the notions of Kendall's τ and correlation coefficient ρ are contrasting but interestingly the sample measure of the former can be expressed as a product moment correlation coefficient. Define $C_{ij} = sgn(X_i - X_j)$ and $D_{ij} = sgn(Y_i - Y_j)$. Assume that $X_i \neq X_j$ and $Y_i \neq Y_j$ for every $i \neq j$ then $\sum_{i,j=1}^{n} C_{ij}^2 = \sum_{i,j=1}^{n} D_{ij}^2 = n(n-1)$. Again $\sum_{i,j=1}^{n} C_{ij} D_{ij} = 2(C - D)$. Thus we can write

$$U = \frac{\sum_{i,j=1}^{n} C_{ij} D_{ij}}{\sqrt{\sum_{i,j=1}^{n} C_{ij}^2 \sum_{i,j=1}^{n} D_{ij}^2}}$$

and hence U is expressed as sample correlation coefficient.

2.5 Summary Measures of U

From the properties of U statistic, $E(U) = \tau$ and

$$Var(U) = \frac{\{2(n-2)\sigma_1^2 + \sigma_2^2\}}{\binom{n}{2}}.$$

Since

$$\phi_1((X_1, Y_1)) = E\{\phi((X_1, Y_1), (X_2, Y_2))|(X_2, Y_2)\},$$

we get

$$\begin{aligned} & E\{sgn(X_1 - X_2)sgn(Y_1 - Y_2)|(X_2, Y_2)\} \\ = & \; 1 - 2F(X_1, \infty) - 2F(\infty, Y_1) + 4F(X_1, Y_1) \end{aligned}$$

Then a routine algebraic manipulation gives

$$\sigma_1^2 = Var\{1 - 2F(X_1, \infty) - 2F(\infty, Y_1) + 4F(X_1, Y_1)\}$$

However,

$$\begin{aligned} \sigma_2^2 &= Var\{sgn(X_1 - X_2)sgn(Y_1 - Y_2)\} \\ &= 1 - \tau^2 \end{aligned}$$

Hence for specific choices of F, the variance can be obtained.

However, if we assume independence, that is $F(x, y) = F(x, \infty)F(\infty, y)\forall(x, y)$, then

$$1 - 2F(X_1, \infty) - 2F(\infty, Y_1) + 4F(X_1, Y_1) \stackrel{D}{=} 4(V - \frac{1}{2})(W - \frac{1}{2}),$$

where V and W are iid R(0,1) variables. Thus we get $\sigma_1^2 = \frac{1}{9}$. Moreover, under independence $\sigma_2^2 = 1$ and hence $Var(U|\text{independence}) = \frac{4n+10}{9n(n-1)}$.

2.6 U is distribution free

Assume independence, that is $F(x, y) = F(x, \infty)F(\infty, y)\forall(x, y)$. Then if F denotes the common univariate DF, then

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \{sgn(X_i - X_j)sgn(Y_i - Y_j)\} \\ &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \{sgn(F(X_i) - F(X_j))sgn(F(Y_i) - F(Y_j))\} \end{aligned}$$

Thus U is expressed in terms of independent R(0,1) variables and hence, under independence, U has a distribution independent of F.

2.7 Symmetry of U

Under independence

$$F(x,y) = F(x,\infty)F(\infty,y)\forall(x,y),$$

we get that for every (i,j),

$$X_i - X_j \stackrel{D}{=} X_j - X_i.$$

Thus $\phi((X_1,Y_1),(X_2,Y_2))$ has the same distribution as $1-\phi((X_1,Y_1),(X_2,Y_2))$ and hence the distributions of C and D are identical. Thus

$$C - D \stackrel{D}{=} D - C$$

and hence

$$U \stackrel{D}{=} -U,$$

that is, under independence, the distribution of U is symmetric about 0.

2.8 A test of independence

Suppose we are interested in testing independence, $H_0 : F(x,y) = F(x,\infty)F(\infty,y)\forall(x,y)$ against dependence that is $H_a : F(x,y) \neq F(x,\infty)F(\infty,y)$ for some (x,y). For underlying dependence, U tends to be either too large or too small. Thus too large and too small values of U indicate the strong possibility of dependence. Again the distribution of U is symmetric about 0 under the null hypothesis. Thus we reject the null hypothesis, if $|T|$ is too large. Again from the theory of U statistic, as $n \to \infty$,

$$\sqrt{n}U \stackrel{D}{\to} N(0,\frac{4}{9}).$$

Then a large sample test rejects H_0 if $\sqrt{\frac{3n}{2}}|U| > \tau_\alpha$.

From the exact distribution of U, one can also calculate the p values to judge the strength of the observed data. Suppose u is the observed value of U. Then the onesided p values are $P_{H_0}(U \geq u)$ and $P_{H_0}(U \leq u)$ corresponding to the underlying alternatives $H_a : F(x,y) \geq$

$F(x,\infty)F(\infty,y)$ for some (x,y) and $H_a: F(x,y) \leq F(x,\infty)F(\infty,y)$ for some (x,y), respectively. However for the two sided alternative $H_a: F(x,y) \neq F(x,\infty)F(\infty,y)$ for some (x,y), the p value is defined as $2min\{P_{H_0}(U \geq u), P_{H_0}(U \leq u)\}$. We reject the null hypothesis if p value does not exceed α in any case.

2.9 Presence of ties

We have assumed continuity of each marginal DF to avoid ties. But in practice, ties can occur within either or both samples. However, U is based the quantities $sgn(X_i - X_j)sgn(Y_i - Y_j)$ and hence assigns a value zero if a tie occurs. Thus as before U remains unbiased for τ even under the presence of ties. Now if we have t ties in X observations of lengths $k_1, .., k_t$ and s ties in Y observations of lengths $k_1^{'}, .., k_s^{'}$, then U can be expressed as

$$U = \frac{\sum_{i,j=1}^{n} C_{ij} D_{ij}}{\sqrt{n(n-1) - \sum_i^t k_i(k_i - 1)}\sqrt{n(n-1) - \sum_{j=1}^{s} k_j^{'}(k_j^{'} - 1)}}$$

The resulting modification is often called Kendall's τ_b measure.

3 Association in discrete distributions

From the development so far, we observe that Kendall's τ is defined only when $p_c + p_d = 1$. In particular if the marginal distributions are not continuous, we have

$$\begin{aligned} \pi_c + p_d &= P\{(X_i - X_j)(Y_i - Y_j) > 0\} + P\{(X_i - X_j)(Y_i - Y_j) < 0\} \\ &= 1 - P\{(Y_i - Y_j)(X_i - X_j) = 0\} \\ &= 1 - P(X_i - X_j = 0 \cup Y_i - Y_j = 0) = 1 - p_0. \end{aligned}$$

Naturally p_0 is the probability that a pair is neither concordant nor discordant. Thus τ is not a sensitive measure of association if $p_0 > 0$.

Thus we define the conditional probabilities of concordance as

$$p_c^* = \frac{P\{(Y_i - Y_j)(X_i - X_j) > 0\}}{P\{(Y_i - Y_j)(X_i - X_j) > 0\} + P\{(Y_i - Y_j)(X_i - X_j) < 0\}}.$$

Naturally $p_c^* = \frac{p_c}{1-p_0}$. Similarly, the conditional probability of discordance is $p_d^* = 1 - p_c^*$. Then a modified association measure can be defined as $\tau^* = p_c^* - p_d^*$. Since $\tau^* = \frac{\tau}{1-p_0}$, we get the corresponding sample version $U^* = \frac{U}{1-\hat{p}_0}$, where $\hat{p}_0$ is a consistent estimate of p_0. Since the large sample distributions of U^* and $U/(1-p_0)$ are the same, the asymptotic distribution and a large sample test based on U^* can be derived as earlier.

4 Application of Kendall's: Test of Trend

For data coming from some measurement processes, often some trend is observed, i.e. observations, in general, depend on time. In the test of randomness, we have seen that a specific pattern of run indicates the presence of trend. Now we shall discuss how Kendall's τ can be used to check the presence of trend.

Suppose X indicates the time variable and Y denotes the corresponding observation. Then an association between X and Y might indicate a presence of trend and this suggests to use Kendall's τ to serve the purpose. Suppose $Y_i = i, i = 1, 2, .., n$ are the time ordered observations, then the hypothesis of independence of the time ordered observations $Y_i, i = 1, 2, .., n$ can be tested through τ. In such a case, we take $X_i = i, i = 1, 2, .., n$ and define the statistic as

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} sgn(j-i) sgn(Y_j - Y_i) \\ &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} sgn(Y_j - Y_i) \end{aligned}$$

U is expected to be higher or lower according as the underlying trend is upward or downward. Thus we reject the null hypothesis of the absence of any trend against the presence of an upward(downward) trend if U is too large(small). Note that the null distribution of U is the same as before and hence tests can be constructed.

Nonparametric Inference: Module 17[1]

What we provide in this module

- Ranks & Midranks
- Properties of ranks
- Linear rank statistics
- Properties

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Ranks & their properties

1.1 Definition of ranks

For N iid observations $Z_i, i = 1, 2, .., N$ from a continuous population the rank of the i th observation is the number of observations that are less than or equal to Z_i. Then R_i, the rank of the i th observation is defined as the number of observations that are less than or equal to the i th observation. That is

$$\begin{aligned} R_i &= \sum_{j=1}^{N} I(Z_j \leq Z_i) \\ &= \#\text{ observations less than or equal to } Z_i. \end{aligned}$$

However, for ordered observations $Z_1 < Z_2 < .. < Z_N$, R_i is simply i. Now observe that

$$\begin{aligned} R_i &= \sum_{j=1}^{N} I(Z_j \leq Z_i) \\ &= N F_N(Z_i), \end{aligned}$$

where F_N is the ecdf. Thus we have the representation

$$R_i = 1 + \sum_{j(\neq i)=1}^{N} I(Z_j \leq Z_i) = 1 + T(Z_i),$$

where the conditional distribution of $T(Z_i)$ is $Binomial(N-1, F(Z_i))$ for fixed Z_i.

1.2 Joint distribution of Ranks

Suppose $\mathbf{Z} = (Z_1, Z_2, .., Z_N)$ are iid observations from an absolutely continuous distribution F, $\mathbf{Y}$ is the corresponding full set of order statistics and $\mathcal{P}_N$ is the set of $N!$ permutations of the first N natural numbers. Then the joint density of $\mathbf{Z}$ given the full set of order statistics is

$$f(\mathbf{z}|\mathbf{y}) = \frac{1}{N!}, \mathbf{z} \in \mathcal{P}_N.$$

Since $\mathbf{Z}$ is fully determined from the knowledge of $\mathbf{Y}$ and the vector of ranks $\mathbf{R} = (R_1, R_2, .., R_N)$, and conversely, we get

$$P(\mathbf{R} = \mathbf{r}|\mathbf{y}) = \frac{1}{N!}, \mathbf{r} \in \mathcal{P}_N.$$

Hence $P(\mathbf{R} = \mathbf{r}) = \frac{1}{N!}, \mathbf{r} \in \mathcal{P}_N$. Thus the joint distribution of $\mathbf{R}$ is independent of any F.

However, one can arrive at the same joint distribution considering the fact that for distinct observations

$$(R_i = r_i \ \forall i = 1, 2, .., N) \Longleftrightarrow (Z_{j1} < Z_{j2} < .. < Z_{jN})$$

for some permutation $(j1, j2, .., jN) \in \mathcal{P}_N$. As a simple consequence of continuity, we get as earlier

$$P(\mathbf{R} = \mathbf{r}) = \frac{1}{N!}, \mathbf{r} \in \mathcal{P}_N.$$

1.3 Marginal distributions of Ranks

Using the definition of marginal distributions, one can obtain the marginal distribution of R_i as

$$\begin{aligned} P(R_i = r_i) &= \sum_{1 \le r_1 \ne .. \ne r_{i-1} \ne r_{i+1} \ne .. \ne r_N \le N} P(\mathbf{R} = \mathbf{r}) \\ &= \frac{1}{N}, r_i = 1, 2, .., N \end{aligned}$$

Hence the marginal distribution of each rank variable is discrete uniform over $\{1, 2, .., N\}$. Thus the following are immediate

$$E(R_i) = \frac{N+1}{2} \quad \text{and} \quad V(R_i) = \frac{N^2 - 1}{12}$$

Now by definition, the marginal joint distribution of R_i and R_j is

$$\begin{aligned} P(R_i = r_i, R_j = r_j) &= \sum_{1 \le r_1 \ne .. \ne r_{i-1} \ne r_{i+1} \ne .. \ne r_{j-1} \ne r_{j+1} \ne .. \ne r_N \le N} P(\mathbf{R} = \mathbf{r}) \\ &= \frac{1}{N(N-1)}, r_i \ne r_j = 1, 2, .., N \end{aligned}$$

As a simple consequence, we have the following $Cov(R_i, R_j) = -\frac{N+1}{12}$ and hence the correlation coefficient between R_i and R_j comes out as $-\frac{1}{N-1}$. Since $\sum_{i=1}^{N} R_i = N(N+1)/2$, R_i and R_j are inversely related and hence the sign of the correlation coefficient become negative. However, for large N, the elements of the rank vector become uncorrelated.

1.4 Midranks: Ranks under tie

However, the distribution of ranks is not so simple when the underlying distribution is discrete. Actually discrete parent allows the possibility of tied observations and hence the usual definition becomes unsuitable. Thus if the underlying distribution is continuous, the two observations being identical has probability zero but for discrete distributions observations can be identical with positive probability. In practice ties can occur due to the limitation of the measurement process and in such a case we introduce the concept of midrank. Midrank is the average rank assigned to each member of a set of tied observations.

We start with an example for better understanding. Suppose the observations are (2,2,7,1,5). Then the sorted observations are (1,2,2,5,7). Then the second and third observations are tied at 2. Thus the midrank assigned to these observations(tied at 2) is (2+3)/2=2.5. Then we get the modified rank vector corresponding to the sorted observations as (1,2.5,2.5,4,5). Thus the midrank of an observation is the number of observations lower than it plus the average rank due to tie. For a set of observations $Z_i, i = 1, 2, .., N$, the midrank of the k th observation is defined as

$$R_k^* = Midrank(Z_k) = \sum_{j=1}^{N} I(Z_j < Z_k) + \frac{\sum_{j=1}^{N} I(Z_j = Z_k) + 1}{2}.$$

Naturally, if the observations are distinct, we get $R_k^* = R_k$ for any $k = 1, 2, .., N$.

For the set of ranks, it was observed that the joint distribution is independent of the parent population. However, in case of midranks such a result does not hold. For example, consider N=2 and assume that Z_i are iid Bernoulli(p). Then the possibilities together with the joint distribution is are given below.

Table 1: **Distribution of midrank for N=2**

(z_1, z_2)	(r_1^*, r_2^*)	$P(R_1^* = r_1^*, R_2^* = r_2^*)$
(0,1)	(1 , 2)	p(1-p)
(1,0)	(2 , 1)	p(1-p)
(1,1)	(1.5 , 1.5)	$(1-p)^2$
(0,0)	(1.5,1.5)	p^2

Unlike the ranks in the untied case, here rank takes three possible values, namely, 1,1.5 and 2 with unequal probabilities. Again it is easy to observe that the marginal distributions are also different from uniform for any $p \in (0, 1)$. As another example, consider three iid Bernoulli(p) variables Z_i. As before, we list the possibilities but obtain only a marginal distribution:

Table 2: **Distribution of midrank for N=3**

(z_1, z_2, z_3)	r_1^*	$P(R_1^* = r_1^*)$
(0,0,0)	2	$(1-p)^3$
(0,0,1)	1.5	$p(1-p)^2$
(0,1,0)	1.5	$p(1-p)^2$
(0,1,1)	1	$(1-p)p^2$
(1,0,0)	2.5	$p(1-p)^2$
(1,0,1)	2.5	$p^2(1-p)$
(1,1,0)	2.5	$p^2(1-p)$
(1,1,1)	2	p^3

Naturally, the marginal distribution is different from uniform.

2 Linear rank statistics

Thus we find that the distribution of ranks is independent of the underlying population as long as the continuity holds. Consequently, ranks form a basis for developing distribution free procedures to identify the possible differences between populations. For example, if one population has a larger location measure, then one can expect higher ranks corresponding to this population in the combined sample observations. Most commonly used tests can be expressed in terms of linear combinations of the function of ranks and indicator of the populations. Such linear combinations are termed linear rank statistics and they can be used for a number of hypothesis testing problems. However, replacing the observations by the corresponding ranks often incurs a loss of information. But it can be shown that the observation and the corresponding rank are highly correlated and hence such loss is not highly consequential.

Suppose $Z_i, i = 1, 2, .., N$ are iid observations from an unknown but continuous distribution F and $R_i, i = 1, 2, ..N$ are the corresponding ranks. Thus it follows from the continuity of F that ranks are distinct with probability one. Then a linear rank statistic is defined by $L = \sum_{i=1}^{N} c_i a(R_i)$, where $c_i, i = 1, 2..$ are regression constants and $a(i), i = 1, 2, ..$ are scores. We have indicated earlier that the joint distribution of ranks is independent of any F. Being a function of ranks, the distribution of L is, therefore, independent of any F. Hence L has the ability to provide distribution free tests for different alternatives regarding F.

2.1 Properties of LRS

Now we shall discuss different exact and asymptotic properties of LRS.

Result :(Summary measures) For a linear rank statistic L,

$$E(L) = N\bar{c}\bar{a},$$

and

$$Var(L) = \frac{N^2}{N-1}\,\sigma_a^2\,\sigma_c^2,$$

where $\bar{c} = \frac{1}{N}\sum\limits_{i=1}^{N} c_i$, $\bar{a} = \frac{1}{N}\sum\limits_{i=1}^{N} a_(i)$, $\sigma_a^2 = \frac{1}{N}\sum\limits_{i=1}^{N}(a_i - \bar{a})^2$ and $\sigma_c^2 = \frac{1}{N}\sum\limits_{i=1}^{N}(c_i - \bar{c})^2$.

Proof: By definition,

$$\begin{aligned} E(L) &= \sum_{i=1}^{N} c_i E\{a(R_i)\} \\ &= \sum_{i=1}^{N} c_i\{\frac{1}{N}a(1) + ... + \frac{1}{N}a(N)\} = N\,\bar{c}\,\bar{a} \end{aligned}$$

Now we note that since the ranks are correlated, different terms of L are, therefore, dependent. Hence $Var(L) = \sum\limits_{i=1}^{N} c_i^2\, V\{a(R_i)\} + \sum\limits_{i\neq j} c_i c_j\, cov(a(R_i), a(R_j))$. Now it is easy to obtain that $Var(a(R_i)) = \sigma_a^2$ and $cov(a(R_i), a(R_j)) = -\frac{\sigma_a^2}{N-1}$. Combining all these the result follows.

Result :(Symmetry) For a linear rank statistic L, if either $c_i + c_{N-i+1}$ or $a(i) + a(N-i+1)$ is a constant for all i, the distribution of L is symmetric about its expectation.

Proof: Observe that $(R_1, .., R_N) \stackrel{D}{=} (N - R_1 + 1, N - R_2 + 1, .., N - R_N + 1)$ and assume that $a(i) + a(N-i+1) = k \forall i$. Then $\bar{a} = k/2$ and hence

$$\begin{aligned} L &\stackrel{D}{=} \sum_{i=1}^{N} c_i a(N - R_i + 1) \\ &\stackrel{D}{=} \sum_{i=1}^{N} c_i(k - a(R_i)) \quad \text{using the assumed condition} \\ &\stackrel{D}{=} kN\bar{c} - L. \end{aligned}$$

Thus using the fact that $k = 2\bar{a}$, we get

$$L - N\bar{c}\bar{a}/2 \stackrel{D}{=} N\bar{c}\bar{a}/2 - L.$$

Hence the symmetry follows. A similar result can be proved using the condition on regression coefficients.

The next result is on asymptotic normality of the linear rank statistic. However, before going to details, we need to define Wald-Wolfowitz, Noether and Hoeffding conditions.

Wald-Wolfowitz Condition(W): A sequence $\{b_n(i), i = 1, 2, .., n\}$ is said to follow Wald-Wolfowitz Condition if

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^r}{\{\frac{1}{n}\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^2\}^{\frac{r}{2}}} = \bigcirc(1)$$

as $n \to \infty$ for $r \geq 3$.

Noether Condition(N): $\{b_n(i)\}$ is said to follow Noether Condition if

$$\frac{\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^r}{\{\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^2\}^{\frac{r}{2}}} = \circ(1)$$

as $n \to \infty$ for $r \geq 3$.

Hoeffding Condition(H): A sequence $\{b_n(i)\}$ is said to satisfy Hoeffding Condition if

$$\max_{0 \leq i \leq n} \frac{(b_n(i) - \bar{b}_n)^2}{\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^2} = \circ(1)$$

as $n \to \infty$.

It can be shown that W condition implies N condition but conditions H and N are equivalent. Now we state the asymptotic normality of linear rank statistic in the following result.

Result:(Asymptotic normality): If

1. $(R_1, .., R_N)$ has a uniform distribution
2. $(c_1, c_2, .., c_N)$ satisfies condition W, and
3. $(a(1), .., a(N))$ satisfies condition N,

then as $N \to \infty$,

$$\frac{L - E(L)}{\sqrt{Var(L)}} \to N(0, 1)$$

.

However, we skip the proof but provide some examples of regression constants and score function satisfying the above requirements.

Suppose $c_i = 0$ or 1 as $i \leq n$ or $i \geq n+1$, where $i = 1, 2, .., N$ with $N = m + n$. Also $m = m(N)$ and $n = n(N)$ are such that $\frac{m}{N} \to \lambda$ and $\frac{n}{N} \to 1 - \lambda$ as $N \to \infty$, where $\lambda \in (0, 1)$. Then it is easy to observe that $(c_1, c_2, .., c_N)$ satisfies W.

On the other hand, if we choose $a(i) = 0$ or 1 as $i \leq [\frac{N+1}{2}]$ or $i \geq [\frac{N+1}{2}] + 1$ with $i = 1, 2, .., N$, then it is easy to obtain that $N^{-1} \max_i (a(i) - \bar{(a)})^2 = \bigcirc(N^{-1})$, as $N \to \infty$. Thus condition H and consequently condition N is satisfied by $\{a(i), i \geq 1\}$.

Nonparametric Inference: Module 18[1]

What we provide in this module

- Applying LRS in two sample location problems
- Applying LRS in two sample scale problems
- Multiple sample location problems

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Two sample problems & LRS

We have already seen that the distribution of LRS is independent of any F under continuity and hence we can use LRS to provide distribution free tests. In addition LRS is asymptotically normal under fairly general conditions and hence large sample tests can also be constructed for various problems. However, we start with two sample location problems.

1.1 Two sample location problem

Suppose

$$X_i \overset{iid}{\sim} F(x), i = 1, 2, .., n$$

independently of

$$Y_j \overset{iid}{\sim} G(x), j = 1, 2, .., m,$$

where F and G are unknown but known to be continuous. For a location problem, we assume $G(x) = F(x - \theta)$ for some $-\infty < \theta < \infty$. Then our objective is to test $H_0 : \theta = 0$ against one of the alternatives $H_a : \theta > 0$ or $H_a : \theta < 0$ or $H_a : \theta \neq 0$. Suppose $\mathbf{Z} = (X_1, X_2, .., X_n, Y_1, .., Y_m)$ is the combined set with $N = m + n$ observations. Also let R_i be the rank of the i th observation in the combined set of observations,i=1,2,..,N. Now we take the population indicators

$$\begin{aligned} c_i &= 0 \text{ if } i \leq n \\ &= 1 \text{ if } i \geq n + 1 \end{aligned}$$

as regression constants. Then LRS reduces to

$$L = \sum_{i=n+1}^{N} a(R_i),$$

which is simply the sum of scores corresponding to the second sample observations. Thus if we can choose an appropriate score, we can develop the corresponding test for the above hypothesis. However, for an appropriate choice of the score function a(.), we need to know

the behaviour of ranks under a location alternative.

Assume $\theta > 0$, then second sample observations are expected to be larger than the first sample observations. Naturally ranks corresponding to the second sample observations are expected to be larger under $\theta > 0$ than under $\theta = 0$. Thus if we take an increasing score function, then $a(R_i)$ are expected to be larger under $\theta > 0$ than under $\theta = 0$ for every $i \geq n+1$. Hence for such a choice, L is expected to be larger under $\theta > 0$ than under $\theta = 0$.

Again for $\theta < 0$, second sample observations are expected to be smaller than the first sample observations. Naturally ranks corresponding to the second sample observations are expected to be smaller under $\theta < 0$ than under $\theta = 0$. Thus for an increasing score function, $a(R_i)$ are expected to be smaller under $\theta < 0$ than under $\theta = 0$ for every $i \geq n + 1$. Hence L is expected to be smaller under $\theta < 0$ than under $\theta = 0$ for such a choice of score.

However for $\theta \neq 0$, second sample observations are expected to be either smaller or larger than the first sample observations. Naturally, in such a situation, ranks corresponding to the second sample observations are expected to be either smaller or larger under $\theta \neq 0$ than under $\theta = 0$. Thus for an increasing score function, $a(R_i)$ are expected to be either smaller or larger under $\theta \neq 0$ than under $\theta = 0$ for every $i \geq n+1$. Hence L is expected to be either smaller or larger under $\theta < 0$ than under $\theta = 0$.

Thus for increasing score function, we get right tailed test based on L for the alternative $H_a : \theta > 0$ and a left tailed test based on L for $H_a : \theta < 0$. For the two sided hypothesis $H_a : \theta \neq 0$, a two tailed test based on L would be appropriate. However, for the selected choice of c_i, the distribution of L under the null hypothesis is never symmetric. But depending on the choice of the score function, the distribution of L can be symmetric about $E_{H_0}(L)$ and hence in such a case the two sided test reduces to a right tailed test based on $|L - E_{H_0}(L)|$. Since the distribution of L is discrete under the null hypothesis, exact size α tests will be, in general, randomized.

1.1.1 Wilcoxon score & test

Consider Wilcoxon score, that is, $a(i) = i$, then L reduces to

$$L = \sum_{i=n+1}^{N} R_i.$$

Thus L is simply the sum of the ranks corresponding to the second sample. L is well known as the Wilcoxon rank sum statistic. It is already derived that $\bar{c} = m/N$ and $\sum_{i=1}^{N}(c_i - \bar(c))^2 = mn/N$. Again for Wilcoxon score $\bar{a} = (N+1)/2$ and $\sigma_a^2 = (N^2-1)/12$. Hence it follows from the theory of LRS that under the null hypothesis, $E_0(L) = m(N+1)/2$ and $Var_0(L) = mn(N+1)/12$.

Again the distribution of L is symmetric under the null hypothesis. First of all we note that for the assumed choice of c_i, we find that $c_i + c_{N-i+1}$ is either 1 or 2. Thus $c_i + c_{N-i+1}$ is not a constant for all i and hence we need to check the other condition. However, for the assumed choice of $a(i)$, we find that $a(i) + a(N-i+1) = N+1 \forall i$ and hence the symmetry of L under the null hypothesis follows from the results of LRS.

It is further interesting to note that Wilcoxon statistic L and Mann Whitney statistic are linearly related. For a proof first of all note that

$$\begin{aligned} \text{Rank of } Y_{(j)} \text{ in the combined sample} &= \text{Number of } X'_{(i)}s \text{ less than } Y_{(j)} \\ &+ \text{Rank of } Y_{(j)} \text{ among the Y observations........}(*) \end{aligned}$$

Now the rank of $Y_{(j)}$ among Y observations is simply j and when summing over j, the first quantity in the RHS of $(*)$ above gives the Mann Whitney statistic U. Again if we sum the LHS of $(*)$ over j, we get Wilcoxon rank sum statistic L. Thus we get the relation

$$L = U + m(m+1)/2.$$

Due to the above relationship, the tests based on Wilcoxon statistic and Mann Whitney statistic are equivalent and hence they share similar properties. It is already derived that tests based on Mann Whitney statistic are consistent. Then due to equivalence, tests based

on Wilcoxon rank sum statistic are also consistent. Further, the asymptotic normality of the Wilcoxon statistic also follow from that of Mann Whitney statistic.

It is easy to observe that Wilcoxon score vector satisfies Noether condition and hence $L^* = \frac{L - E_{H_0}(L)}{\sqrt{Var_{H_0}(L)}}$ is asymptotically standard normal. Thus we can perform large sample tests based on L^*. In particular, a large sample test for $H_0 : \theta = 0$ against $H_a : \theta > 0$ rejects the null hypothesis if $L^* > \tau_\alpha$. Similarly large sample tests for the remaining alternatives can also be constructed.

1.1.2 Median score & test

Next we choose $a(i) = I(i > [\frac{N+1}{2}])$. Since $[\frac{N+1}{2}]$ is the median of the combined set $\{1, 2, .., N\}$, we have score unity when i exceeds the median and zero otherwise. This justifies the name median score. For median score, L reduces to

$$L = \sum_{i=n+1}^{N} I(R_i > [\frac{N+1}{2}]),$$

which gives the number of the second sample ranks exceeding $[\frac{N+1}{2}]$. L is well known as Mood's median test statistic. For Median score $\bar{a} = \frac{N-[(N+1)/2])}{N}$ and $\sigma_a^2 = \frac{N-[(N+1)/2])}{N}(1 - \{\frac{N-[(N+1)/2])}{N}\})$. Now from the first principle, it follows that

$$\begin{aligned} P_{H_0}(L = l) &= \frac{\#\text{ choices of } (R_{n+1}, .., R_N) \text{ out of } (1, 2, .., N) : L = l}{\binom{N}{n}} \\ &= \frac{\binom{N-[(N+1)/2]}{l}\binom{[(N+1)/2]}{m-l}}{\binom{N}{n}} \end{aligned}$$

for admissible choices of l. Thus the distribution is hypergeometric.Now it follows easily from the theory of LRS that under the null hypothesis, $E_0(L) = m(1 - [(N+1)/2]/N)$. In a similar way, using the properties of hypergeometric distribution, $Var_0(L)$ can also be obtained. Now for the median score, we find that $a(i) + a(N - i + 1) = 1 \forall i$. Thus the symmetry of L under the null hypothesis follows from the results of LRS.

From the general principle, the exact tests can be constructed. However, the median score vector satisfies Hoeffding condition and hence $L^* = \frac{L - E_{H_0}(L)}{\sqrt{Var_{H_0}(L)}}$ is asymptotically standard

normal. Thus we can construct different large sample tests based on L^*. In particular, a large sample test for $H_0 : \theta = 0$ against $H_a : \theta \neq 0$ rejects the null hypothesis if $|L^*| > \tau_{\alpha/2}$. Similarly large sample tests for the remaining alternatives can also be constructed.

1.2 LRS in scale problems

Suppose

$$X_i \overset{iid}{\sim} F(\frac{x-\mu_1}{\sigma_1}), i = 1, 2, .., n$$

independently of

$$Y_j \overset{iid}{\sim} F(\frac{x-\mu_2}{\sigma_2}), j = 1, 2, .., m,$$

where F is unknown but known to be continuous. If $\sigma_1 = \sigma_2$ but $\mu_1 \neq \mu_2$, the distribution of rank vector is not uniform. Hence no exactly or even asymptotically distribution free test can be constructed using L. But, if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$, the distribution of rank vector is uniform and hence exactly distribution free tests can be derived using L. However, if $\sigma_1 = \sigma_2$ and $\mu_1 \neq \mu_2$ with (μ_1, μ_2) known, exactly distribution free test can also be derived using L. Thus in order to get distribution free tests using L, we assume $X_i \overset{iid}{\sim} F(x), i = 1, 2, .., n$ independently of $Y_j \overset{iid}{\sim} F(\frac{x}{\sigma}), j = 1, 2, .., m$, with $\sigma > 0$. That is, we assume that the two distributions have a common location and they differ only in scale. Our objective is to test $H_0 : \sigma = 1$ against all alternatives. As earlier we take the LRS as $L = \sum_{i=n+1}^{N} a(R_i)$, that is L becomes the sum of scores corresponding to the second sample observations.

As earlier, we need to enumerate the behaviour of ranks under a scale alternative to decide an appropriate choice of the score function a(.).

First of all assume that the observations are positive with probability one, that is, $F(0) = 0$. Then

$$\sigma > 1 \Longleftrightarrow F(x) \geq F(\frac{x}{\sigma}) \forall x \Longleftrightarrow Y \overset{st}{>} X$$

and

$$\sigma < 1 \Longleftrightarrow F(x) \leq F(\frac{x}{\sigma}) \forall x \Longleftrightarrow Y \overset{st}{<} X.$$

Thus tests for stochastic alternatives(e.g. Mann-Whitney test) are all applicable for testing $H_0 : \sigma = 1$ against all alternatives.

Next assume observations on the whole real line together with symmetry, that is $F(x) + F(-x) = 1 \forall x$. Then $\sigma > 1$ implies that the second sample observations are expected to have lower concentration than the first sample observations. Naturally observations corresponding to the second sample observations are expected to be either too large or too small. Thus second sample ranks are expected to be either too smaller or too larger under $\sigma > 1$ than under $\sigma = 1$. Thus if we take a symmetric U shaped score function, then $a(R_i)$ are expected to be larger under $\sigma > 1$ than under $\sigma = 1$ for every $i \geq n + 1$. Hence L is expected to be larger under $\sigma > 1$ than under $\sigma = 1$ and hence a right tailed test is suggested. Similarly L is expected to be smaller under $\sigma < 1$ than under $\sigma = 1$ and hence a left tailed test is suggested. Again for testing $\sigma = 1$ against $\sigma \neq 1$ equal tailed test based on $|L - E_{H_0}(L)|$ is suggested provided the distribution of L is symmetric under the null hypothesis. One can also use a bell shaped symmetric score function, but in such a case, the rejection regions will be reversed. However, the distribution of L is discrete under the null hypothesis and hence exact size α tests will be, in general, randomized.

1.2.1 Quartile score & test

Note that for the set $\{1, 2, .., N\}$, the first and third quartiles are $[\frac{N}{4}]$ and $[\frac{3N}{4}]$, respectively. In quartile score, we assign unity if i lies outside the interval $([\frac{N}{4}], [\frac{3N}{4}])$ and assign zero otherwise. This suggests to define $a(i) = 1 - I([\frac{N}{4}] < i < [\frac{3N}{4}])$ where $i = 1, 2, .., N$. Then using the first principle, it is easy to obtain that

$$P_{H_0}(L = l) \quad = \quad \frac{\binom{2[N/4]}{l}\binom{N-2[N/4]}{m-l}}{\binom{N}{n}}$$

for admissible choices of l. As earlier, the large sample tests can be performed as usual.

1.2.2 Mood's score, Ansari-Bradley & Klotz normal score

Mood, Ansari-Bradley and Klotz normal scores correspond to the choices $a(i) = (i - \frac{N+1}{2})^2$, $a(i) = |i - \frac{N+1}{2}|$ and $a(i) = \{\Phi^{-1}(\frac{i}{N+1})\}^2$. As earlier summary measures and large sample tests based on the standardised statistics can be constructed.

2 Multiple sample location problem

We have discussed so far tests for two independent samples. But un practice, we can have more that two populations to compare. Nonparametric methods can also be developed in such a situation. Specifically, we assume that

$$X_{ij} \overset{independently}{\sim} F(x - \theta_i), j = 1, 2, .., n_i, i = 1, 2, .., k,$$

for unknown but continuous F. Then our objective is to test $H_0 : \theta_1 = \theta_2 = .. = \theta_k$ against all alternatives. Thus we get a nonparametric analogue of one way classified data.

2.1 An intuitive test

Define R_i as the rank of the i th observation in the combined sample $(X_{11}, .., X_{1n_1}, .., X_{kn_k})$ of $N = \sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i$ observations. Then the distribution of ranks $R_i, i = 1, 2, .., N$ is uniform under the null hypothesis. If the null hypothesis is true then the total sum of ranks, that is, $N(N + 1)/2$ is expected to be divided proportionally among the k samples. That is under H_0 , the sum of ranks for the i th sample is expected to be $n_i(N + 1)/2$. If the sum of the observed ranks assigned to the ith sample is denoted by R_{i0}, a reasonable test statistic could be based on the deviations between these observed and expected rank sums. Thus one can use the sum of these squared deviations as a reasonable statistic. In particular, we suggest

$$S = \sum_{i=1}^{k} \{R_{i0} - n_i(N + 1)/2\}^2$$

to measure the deviation from the null hypothesis. Naturally a large value of S indicates deviation from the null and hence we reject the null hypothesis if S is too large.

2.2 Kruskal-Wallis Test

Quite naturally S provides a distribution free test for the null hypothesis. However, a more useful test criterion is a weighted sum of squares of deviations, where the sample size reciprocals are the weights. Following the above argument, Kruskal and Wallis (1952) defined the statistic

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} {n_i}^{-1} \{R_{i0} - n_i(N+1)/2\}^2.$$

As before, a large value of H indicates deviation from the null and hence we reject the null hypothesis if H is too large. However tests based on S and H are equivalent, when the number of observations from each sample are the same.

A computationally easier form of H can be expressed as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_{i0}^2}{n_i} - 3(N+1).$$

As before, being a function of ranks, H provides an exactly distribution free test. In addition, if $\min(n_1, .., n_k) \to \infty$ such that $\frac{n_i}{N} \to \lambda_i \in (0,1)$ then it can be proved that $H \xrightarrow{D} \chi^2_{k-1}$ under the null hypothesis. Thus an asymptotically size α test rejects the null hypothesis if $H > \chi^2_{k-1,\alpha}$. However, for more exposure on multi-sample problems, we refer the reader to the book by Gibbons and Chakraborti(2004).

Nonparametric Inference: Module 19[1]

What we provide in this module

- Review of the concept of efficiency in estimation theory.
- Comparison of tests
- Finite sample relative efficiency: The concept and problems
- Asymptotic power: Definition

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 A review of performance comparison in estimation

1.1 Small sample comparison

In statistical point estimation, we compare two estimators through some expected distance measure from the parameter of interest. For example, suppose we have two estimators, T_{1n} and T_{2n} based on n observations for the estimation of some real valued parameter θ. Then the relative efficiency can be measured by the ratio of Mean Square Errors(MSE's)

$$\frac{E(T_{1n}-\theta)^2}{E(T_{2n}-\theta)^2},$$

provided the existence. Naturally, the estimator with the smallest MSE is preferable. If the estimators are unbiased, the above ratio reduces to the ratio of variances or the ratio of precisions. Then, in such a situation, estimator with the smallest variance is the best. However, such a criterion depends heavily on the choice of the distance measure in addition to the existence of second order moments. Thus we need another criterion, known as Pitman Measure of Closeness(PMC), defined by

$$\mathcal{P}(T_{1n}, T_{2n}|\theta) = P(|T_{1n}-\theta| < |T_{2n}-\theta|).$$

Then, in its simplest form, PMC measures the relative frequency with which T_{1n} is closer than T_{2n} from the unknown parameter of interest θ. However, PMC does not measure the relative distances of T_{1n} and T_{2n} from the true θ. But, calculation of PMC requires the sampling distribution of T_{1n} and T_{2n} and hence is not easy to compute for a number of estimators.

1.2 Large sample comparison

Now assume that we have two consistent estimators T_{1n} and T_{2n} of some real valued parameter θ, that is $T_{kn} \xrightarrow{P} \theta$ for each k=1,2. Since consistency is a desirable property in any estimation problem, rate of convergence would be a natural measure of relative performance. Thus, the estimator converging to θ in a higher rate is better. But the problem is that rate of convergence is not easy to compute theoretically. However, if we assume the existence of MSE, then $MSE(T_{kn}) = E(T_{kn} - \theta)^2 \to 0$ as $n \to \infty$ ensures consistency of $T_{kn}, k = 1, 2$. Thus the rate of convergence can be judged from the ordering of the MSE's. That is, the estimator corresponding to smaller MSE converges at a faster rate to θ in probability than its competitor. In other words, if $MSE(T_{1n}) < MSE(T_{2n}) \forall \theta$, T_{1n} is better in terms of the convergence rate. But again, such a comparison involves huge calculation and hence lacks practical feasibility.

Thus we restrict to the class of estimators, which are not only consistent but also asymptotically normal. That is for every $k = 1, 2$

$$\sqrt{n}(T_{kn} - \theta) \xrightarrow{D} N(0, \sigma_k^2)$$

for some σ_k^2. Then as before, relative efficiency can be measured by the ratio of asymptotic variances. However, in this case the efficiency is called the Asymptotic Relative Efficiency(ARE). In brief, suppose

$$\sqrt{n}(T_{1n} - \theta) \xrightarrow{D} N(0, \tau^2)$$

and

$$\sqrt{n}(T_{2n'} - \theta) \xrightarrow{D} N(0, \tau^2),$$

where $T_{2n'}$ is based on $n' = n'(n)$ observations. Then ARE of $\{T_{1n}\}$ with respect to $\{T_{2n}\}$ is defined as

$$e_{12} = \lim_{n \to \infty} \frac{n'(n)}{n}$$

provided the limit exists and is independent of the subsequence n'. Thus ARE can be interpreted as the relative number of additional observations required by the less efficient estimator to reach the same accuracy. Naturally, $\{T_{1n}\}$ is more(or less) efficient in the asymptotic sense if $e_{12} > 1$(or < 1). However, in actual practice, the limit is not easy to evaluate and we use the equivalent form

$$e_{12} = \frac{\sigma_2^2}{\sigma_1^2}.$$

2 Performance comparison in hypothesis testing

Suppose we have two competing tests for the same hypothesis. Then it is natural to inquire about the relative performance. In estimation theory, we have used some distance measure to measure the performance. In hypothesis testing problems, the only relevant measure is power, that is, a bounded measure of departure from the value specified by the null hypothesis. Thus if the tests are of same level of significance, then the test having a power curve lying above the power curves of all the competitors would be preferred. However, in nonparametric inference, the underlying distributions are never specified and hence construction of optimum tests is almost impossible. Therefore, at a minimum, we can compare the power functions for fixed sample size. But such a comparison depends on the sample size, the alternative, the form of the underlying distribution and the level of significance. Hence reaching a general conclusion will be difficult.

Therefore, as an alternative one can use the the large sample methods to compare the tests. Although use of large sample approximation eliminates the dependence on sample size but creates a more serious problem. Because, we, in general, consider consistent tests so that power reaches unity for large sample size. Thus the limiting power function is of no use for the purpose of comparison. Pitman(1948), in this context suggested to compare asymptotic values of power considering sequence of

alternatives converging to the value provided by the null hypothesis. An appropriate choice of the sequence of alternatives enables a clear comparison.

2.1 Finite sample efficiency

Suppose $X_1, X_2, .., X_n$ are iid observations from an unknown distribution G and our interest lies in testing $H_0 : G \in \Omega_0$ against $H_a : G \in \Omega_a$, where $\Omega_0(\Omega_a)$ is the class of distributions specified by the null(alternative) hypothesis.
Consider two different level α tests, $\phi_n^{(1)}$ and $\phi_n^{(2)}$. That is, for any k=1,2,

$$E_G \phi_n^{(k)} \leq \alpha \text{ for every } G \in \Omega_0$$

Take an arbitrary $G^* \in \Omega_a$ and fixed β such that $\alpha < \beta < 1$. If N_k denotes the minimum number of observations required by test k to exceed a prefixed power β, Then the relative efficiency of test 1 with respect to test 2 is the ratio of the required sample sizes to attain power atleast β. Precisely, if N_k denotes the minimum number of observations required by test k to satisfy

$$E_G^* \phi_{N_k}^{(k)} \geq \beta \text{ for any } k = 1, 2,$$

then the fixed sample relative efficiency is defined by the ratio

$$e_{12}(\alpha, \beta, G^*) = \frac{N_B(\alpha, \beta, G^*)}{N_A(\alpha, \beta, G^*)}.$$

Naturally test 1 is preferred over test 2 if $e_{12}(\alpha, \beta, G^*) > 1$ and if the opposite holds, we prefer test 2 to test 1.

2.2 Asymptotic Power(AP)

Now $e_{12}(\alpha, \beta, G^*)$ depends on α, β and G^* and hence, as discussed earlier, makes the comparison difficult. Thus, before defining the asymptotic relative efficiency, we discuss the concepts of local alternatives and asymptotic power. Assume that the

underlying distributions can be indexed by a real parameter θ so that the problem reduces to testing $H_0 : \theta \in \Omega_0$ against $H_a : \theta \in \Omega_a$. Then a sequence of alternatives $\{\theta_n\}$ is called Pitman's local alternative if

(i) $\theta_n \in \Omega_a \quad$ for all $\; n \geq 1$, and

(ii) $\theta_n \to \theta_0 \in \Omega_0 \;$ as $\; n \to \infty$

Now consider a sequence of tests $\{\phi_n\}$ for testing $H_0 : \theta \in \Omega_0$ against $H_a : \theta \in \Omega_a$. Then $\{\phi_n\}$ is asymptotically size α if

$$E_\theta \phi_n \to \alpha \quad \text{for all} \;\; \theta \in \Omega_0$$

Then the AP of $\{\phi_n\}$ against the local alternatives $\{\theta_n\}$ is defined as $\lim_{n\to\infty} E_{\theta_n}\phi_n$ provided the limit exists and lies in $(\alpha, 1)$.

However, determination of AP is not straightforward and hence we provide below an equivalent expression considering some assumptions.

Consider a sequence of asymptotically size α right tailed tests $\{\phi_n\}$ based on $\{T_n\}$ for testing $H_0 : \theta \in \Omega_0$ against $H_a : \theta \in \Omega_a$. That is

$$\begin{aligned} \phi_n &= 1 \text{ if } T_n > c_n \\ &= 0 \text{ if } T_n \leq c_n \end{aligned}$$

where $\{c_n\}$ are such that

$$E_\theta \phi_n \to \alpha \quad \text{for all} \;\; \theta \in \Omega_0.$$

Then the AP of $\{\phi_n\}$ against $\{\theta_n\}$ is given by $\beta = \lim_{n\to\infty} E_{\theta_n}\phi_n \in (\alpha, 1)$, provided the limit exists.

Nonparametric Inference: Module 20[1]

What we provide in this module

- Asymptotic Power(AP)-The determination
- Asymptotic relative efficiency: Definition and working formula
- Few illustrations with interpretation

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Asymptotic Power(AP)-The determination

For the determination of AP, we need the following regularity conditions:

A1. Suppose $\theta_0 \in \Omega_0$ and $\theta \in \Omega_a$ are such that $\theta > \theta_0$. Then we take $\theta_n = \theta_0 + \frac{b}{\sqrt{n}}, b > 0$ so that $\theta_n \in \Omega_a \forall n$ and $\theta_n \to \theta_0 \in \Omega_0$ as $n \to \infty$.

A2. Suppose there exists $\mu_{T_n}(\theta)$ and $\sigma_{T_n}(\theta) > 0$ such that under θ_0

$$\frac{T_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)} \xrightarrow{D} W,$$

and under $\{\theta_n\}$,

$$\frac{T_n - \mu_{T_n}(\theta_n)}{\sigma_{T_n}(\theta_n)} \xrightarrow{D} W$$

for some known continuous random variable W with distribution function $H(w) = P(W \leq w)$.

A3. $\frac{d\mu_{T_n}(\theta)}{d\theta}$ exists and is non zero and continuous in the neighbourhood of θ_0.

A4. There exists a positive constant d such that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\mu'_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)} = d,$$

with $\mu'_{T_n}(\theta_0) = \frac{d\mu_{T_n}(\theta)}{d\theta}|_{\theta_0}$

A5. $\lim_{n\to\infty} \frac{\sigma_{T_n}(\theta_n)}{\sigma_{T_n}(\theta_0)} = 1.$

Result 1: *Under the above conditions, the asymptotic power(AP) of a test is $P(W > w_\alpha - bd)$, where w_α is such that $P(W > w_\alpha) = \alpha$.*

Proof: Since $\{\phi_n\}$ is asymptotically size α, we get

$$\begin{aligned} \alpha &= \lim_{n\to\infty} P_{\theta_0}(T_n > c_n) \\ &= \lim_{n\to\infty} P_{\theta_0}(\frac{T_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)} > \frac{c_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)}) \\ &= P(W > w_\alpha) \text{ using A.2,} \end{aligned}$$

where w_α is such that $H(w_\alpha) = 1 - \alpha$. Thus, we get

$$\frac{c_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)} \to w_\alpha \qquad\qquad (*)$$

as $n \to \infty$. Now AP of $\{\phi_n\}$ against $\{\theta_n\}$ is given by

$$\begin{aligned} AP &= \lim_{n\to\infty} P_{\theta_n}(T_n > c_n) \\ &= \lim_{n\to\infty} P_{\theta_n}\left(\frac{T_n - \mu_{T_n}(\theta_n)}{\sigma_{T_n}(\theta_n)} > \frac{c_n - \mu_{T_n}(\theta_n)}{\sigma_{T_n}(\theta_n)}\right) \end{aligned}$$

Since

$$\begin{aligned} \frac{c_n - \mu_{T_n}(\theta_n)}{\sigma_{T_n}(\theta_n)} &= \frac{c_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)}\frac{\sigma_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_n)} - (\theta_n - \theta_0)\frac{\mu_{T_n}(\theta_n) - \mu_{T_n}(\theta_0)}{(\theta_n - \theta_0)\sigma_{T_n}(\theta_0)}\frac{\sigma_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_n)} \\ &= \frac{c_n - \mu_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)}\frac{\sigma_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_n)} - \frac{b}{\sqrt{n}\sigma_{T_n}(\theta_0)}\frac{\mu_{T_n}(\theta_n) - \mu_{T_n}(\theta_0)}{(\theta_n - \theta_0)}\frac{\sigma_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_n)} \\ &\to w_\alpha - bd \text{ as } n \to \infty, \text{ (using A1,A3-A5 and (*))} \end{aligned}$$

we have $AP = P(W > w_\alpha - bd)$. Hence the proof.

It is worth mentioning that $AP \in (\alpha, 1)$.

2 Asymptotic Relative Efficiency(ARE)

2.1 The definition

Now consider the testing problem $H_0 : \theta \in \Omega_0$ against $H_a : \theta \in \Omega_a$. Suppose we have two tests $\phi_n^{(1)}$ based on $\{T_{1n}\}$ and $\phi_n^{(2)}$ based on $\{T_{2n}\}$. Both the tests are assumed to be right tailed and asymptotically size α. Let $\{k(n)\}$ and $\{k'(n)\}$ be two sequence of positive integers such that

(i) $\{k(n)\}$ and $\{k'(n)\}$ are both increasing in n and

(ii) as $n \to \infty$, $k(n) \to \infty$ and $k'(n) \to \infty$.

Again these sequences are such that

$$\alpha < \lim_{n\to\infty} P_{\theta_n}(T_{1k(n)} > c_{1n}) = \lim_{n\to\infty} P_{\theta_n}(T_{2k'(n)} > c_{2n}) < 1.$$

Then ARE of $\phi_n^{(1)}$ relative to and $\phi_n^{(2)}$ is defined by

$$ARE(1|2) = \lim_{n\to\infty} \frac{k'(n)}{k(n)}$$

provided the limit exists and is independent of the sequences $\{k(n)\}$ and $\{k'(n)\}$ satisfying (i) and (ii) above.

Thus ARE is the limit of the ratio of the sample sizes required to reach the same limiting power against the identical sequence of alternatives for tests with the same limiting significance levels. Then $ARE(1|2) = 2$ means that we need approximately twice as many observations for test 2 as compared to test 1 to reach the same limiting power for the same sequence of alternatives. Naturally test 1 is better or worse than test 2 in the limit according as $ARE(1|2) > 1$ or < 1.

2.2 The determination

However, the definition of ARE as given above is not enough to compute in practice. Thus we need some equivalent expression so that the efficiency can be evaluated easily. The following result gives an equivalent expression of ARE.

Result 2: *If the tests $\phi_n^{(1)}$ and $\phi_n^{(2)}$ satisfy the regularity conditions (A1)-(A5), then*

$$ARE(1|2) = \frac{\lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{1n}}(\theta_0)}{\sigma_{T_{1n}}(\theta_0)}\}^2}{\lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{2n}}(\theta_0)}{\sigma_{T_{2n}}(\theta_0)}\}^2}.$$

Proof: From the definition, the sequences $\{k(n)\}$ and $\{k'(n)\}$ are such that

$$\lim_{n\to\infty} P_{\theta_n}(T_{1k(n)} > c_{1k(n)}) = \lim_{n\to\infty} P_{\theta_n}(T_{2k'(n)} > c_{2k'(n)}).$$

Since, $\phi_n^{(1)}$ and $\phi_n^{(2)}$ satisfy the regularity conditions (A1)-(A5), $\lim_{n\to\infty} P_{\theta_n}(T_{1k(n)} > c_{1k(n)})$ is nothing but the AP of $\phi_n^{(1)}$. Then we have from Result 1 that

$$\lim_{n\to\infty} P_{\theta_n}(T_{1k(n)} > c_{1k(n)}) = P(W > w_\alpha - bd_1).$$

Similarly, for $\phi_n^{(2)}$ we get the expression of AP as

$$\lim_{n\to\infty} P_{\theta_n}(T_{2k'nn} > c_{2k'(n)}) = P(W > w_\alpha - bd_2).$$

Thus we require $\{k(n)\}$ and $\{k'(n)\}$ satisfying

$$\begin{aligned}
\lim_{n\to\infty} P_{\theta_n}(T_{1k(n)} > c_{1k(n)}) &= \lim_{n\to\infty} P_{\theta_n}(T_{2k'(n)} > c_{2k'(n)}) \\
\Leftrightarrow P(W > w_\alpha - bd_1) &= P(W > w_\alpha - bd_2) \\
\Leftrightarrow \lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{1k(n)}}(\theta_0)}{\sigma_{T_{1k(n)}}(\theta_0)} &= \lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{2k'(n)}}(\theta_0)}{\sigma_{T_{2k'(n)}}(\theta_0)} \\
\Leftrightarrow \lim_{n\to\infty}\sqrt{\frac{k(n)}{n}}\lim_{n\to\infty}\{\frac{1}{\sqrt{k(n)}}\frac{\mu'_{T_{1k(n)}}(\theta_0)}{\sigma_{T_{1k(n)}}(\theta_0)}\} &= \lim_{n\to\infty}\sqrt{\frac{k'(n)}{n}}\lim_{n\to\infty}\{\frac{1}{\sqrt{k'(n)}}\frac{\mu'_{T_{2k'(n)}}(\theta_0)}{\sigma_{T_{2k'(n)}}(\theta_0)}\} \\
\Leftrightarrow \lim_{n\to\infty}\frac{k'(n)}{k(n)} &= \frac{\lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{1n}}(\theta_0)}{\sigma_{T_{1n}}(\theta_0)}\}^2}{\lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{2n}}(\theta_0)}{\sigma_{T_{2n}}(\theta_0)}\}^2}.
\end{aligned}$$

This completes the proof.

Remark1: From the expression of ARE, we observe that under the regularity conditions, the ARE can be interpreted as the ratio of two efficacy measures, where efficacy of test based on T_n is defined as $\lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_n}(\theta_0)}{\sigma_{T_n}(\theta_0)}\}^2$.
Remark2: The expression of ARE remains valid even for two sided alternatives.
Remark3: The ARE does not depend on α and β as long as the regularity conditions are satisfied. Thus unlike fixed sample efficiency, the ARE does not suffer the disadvantages.

2.3 Examples

Now we shall provide some examples on finding the ARE in some testing problems. Suppose $X_i, i = 1, 2, .., n$ are iid observations from $F(x-\theta), -\infty < \theta < \infty$, where $F(x) + F(-x) = 1 \forall x$. Consider testing $H_0 : \theta = 0$ against $H_a : \theta > 0$. Now we have the following tests:

- Test 1: Consider a test based on the t statistic $T_{1n} = \sqrt{n}\frac{\bar{X}}{s_n}$, which rejects the null hypothesis if $T_{1n} > c_{1n}$, where $(n-1)s_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2$.

- Test 2: Consider Sign test based on the statistic $T_{2n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i > 0)$, which rejects the null hypothesis if $T_{2n} > c_{2n}$

- Test 3: Consider Wilcoxon signed rank test based on $T_{3n} = \frac{\sum_{i\le j} I(\frac{X_i+X_j}{2}>0)}{\binom{n}{2}}$, which rejects the null hypothesis if $T_{3n} > c_{3n}$.

Now consider the sequence of local alternatives $\theta_n = \frac{b}{\sqrt{n}}$ and assume $\sigma^2 = Var_F(X_1) < \infty$. First of all , we verify the conditions (A2)-(A5). Since $s_n \xrightarrow{P} \sigma$, $\sqrt{n}\frac{\bar{X}}{s_n}$ and $\sqrt{n}\frac{\bar{X}}{\sigma}$ have the same asymptotic distribution. This suggests to define $\mu_{T_{1n}}(\theta) = \sqrt{n}\frac{\theta}{\sigma}$ and $\sigma_{T_{1n}}(\theta) = 1$. Then conditions (A3)-(A5) are all satisfied with $d_1 = \sigma^{-1}$. Now applying Central Limit Theorem(CLT), it can be shown that

$$\frac{T_{1n} - \mu_{T_{1n}}(\theta)}{\sigma_{T_{1n}}(\theta)} \xrightarrow{D} N(0,1),$$

under both $\theta = 0$ and $\theta = \theta_n$. Thus (A2) is also satisfied. Hence the efficacy of Test 1 is

$$\begin{aligned} e(T_{1n}) &= \lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{1n}}(0)}{\sigma_{T_{1n}}(0)}\}^2 \\ &= \{\frac{\sqrt{n}/\sigma}{\sqrt{n}}\}^2 = \sigma^{-2} \end{aligned}$$

Next consider Sign test statistic. Then $E_\theta(T_{2n}) = P_\theta(X_1 > 0) = F(\theta)$,$\sigma_{T_{2n}}(\theta) = \sqrt{\frac{F(\theta)(1-F(\theta))}{n}}$. Since, by DeMoivre-Laplace limit theorem

$$\frac{T_{2n} - F(\theta)}{\sqrt{\frac{F(\theta)(1-F(\theta))}{n}}} \xrightarrow{D} N(0,1),$$

under both $\theta = 0$ and $\theta = \theta_n$, we define $\mu_{T_{2n}}(\theta) = F(\theta)$ and $\sigma_{T_{2n}}(\theta) = \sqrt{\frac{F(\theta)(1-F(\theta))}{n}}$. Thus conditions (A2)-(A5) are all satisfied with $d_2 = 2f(0)$, where $f(x) = \frac{dF(x)}{dx}$ exists at all x and continuous with $f(0) > 0$. Hence the efficacy of Test 2 is

$$\begin{aligned} e(T_{2n}) &= \lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{2n}}(0)}{\sigma_{T_{2n}}(0)}\}^2 \\ &= \{\frac{f(0)}{\sqrt{n}\sqrt{\frac{1}{4n}}}\}^2 = 4f^2(0). \end{aligned}$$

Now consider Wilcoxon Signed rank test statistic. Then T_{3n} can be written as

$$
\begin{aligned}
T_{3n} &= \frac{1}{\binom{n}{2}}\{\sum_{i=1}^{n} I(X_i > 0) + \sum_{i<j} I(X_i + X_j > 0)\} \\
&= \frac{2}{n-1}\frac{1}{n}\sum_{i=1}^{n} I(X_i > 0) + \frac{1}{\binom{n}{2}}\sum_{i<j} I(X_i + X_j > 0) \\
&= \frac{2}{n-1}U_1 + U_2,
\end{aligned}
$$

where U_1 and U_2 are U statistics corresponding to the kernels $I(X_1 > 0)$ and $I(X_1+X_2 > 0)$, respectively. Then from the theory of U statistics, it follows that $T_{3n} - U_2 \xrightarrow{P} 0$ as $n \to \infty$. That is T_{3n} and U_2 have the same asymptotic distribution. Now from the theory of U statistics it follows that as $n \to \infty$,

$$
\frac{U_2 - E_\theta(U_2)}{\sqrt{Var_\theta(U_2)}} \xrightarrow{D} N(0, 1),
$$

under both $\theta = 0$ and $\theta = \theta_n$. Thus we set $\mu_{T_{3n}}\theta = E_\theta(U_2) = P_\theta(X_1 + X_2 > 0) = \int F(x + 2\theta)dF(x)$ and $\sigma_{T_{3n}}\theta = \sqrt{Var_\theta(U_2)} = \sqrt{\frac{4Var_\theta F(X_1+\theta)}{n}}$. Thus conditions (A2)-(A5) are all satisfied. Now assume that $f(x) = \frac{dF(x)}{dx}$ exists and is continuous at all x. In addition, assume that differentiation under the sign of integration is valid. Then we get $\mu'_{T_{3n}}(0) = 2\int_{-\infty}^{\infty} f^2(x)dx$ assuming $\int_{-\infty}^{\infty} f^2(x)dx < \infty$. Again, under $\theta = 0$, $F(X_1)$ has a $R(0, 1)$ distribution and hence we get $\sigma_{T_{3n}}(0) = \frac{1}{\sqrt{3n}}$. Thus we get the efficacy of Test 3 as

$$
\begin{aligned}
e(T_{3n}) &= \lim_{n\to\infty}\{\frac{1}{\sqrt{n}}\frac{\mu'_{T_{3n}}(0)}{\sigma_{T_{3n}}(0)}\}^2 \\
&= \{2\sqrt{3}\int_{-\infty}^{\infty} f^2(x)dx\}^2
\end{aligned}
$$

Now we can compare different pairs of tests using ARE. First of all consider comparing Sign test and test based on t statistic. The ARE takes the form

$$
\begin{aligned}
ARE(Sign|t) &= \frac{e(T_{2n})}{e(T_{1n})} \\
&= 4\sigma^2 f^2(0)
\end{aligned}
$$

Naturally, ARE depends on the underlying density f. Thus we provide below the expressions corresponding to different distributions. Thus we observe that Sign test is always less efficient

Table 1: **Comparing Sign Test and test based on t statistic**

$f(x)$	$ARE(Sign\|t)$
N(0,1)	$\frac{2}{\pi} = .64$
Logistic(0,1)	$\frac{\pi^2}{12} = .82$
Double Exponential(0,1)	2.0
R(-1,1)	$\frac{1}{3\pi} = .33$
Cauchy(0,1)	∞

than the t statistic based test except for Double exponential distribution. Roughly speaking, for normal parents, we would require 36% less observations for t test compared to Sign test to get the same performance. This is not unexpected as the sign test uses only the information about the signs of the differences. However for heavy tailed distribution like Double exponential, the quality of information used by the sign test is improved and hence becomes twice as efficient as the test based on t statistic. This shows that t statistic based test is no longer desired for double exponential distributions. For information, we note that within the class of all continuous unimodal symmetric densities, $ARE(Sign|t)$ can not be lower than $\frac{1}{3}$(see, Hodges and Lehmann,1956, for details). Form the above table, we find that the only distribution attaining the lower bound is $R(-1,1)$.

Now consider comparing Wilcoxon Signed rank test and test based on t statistic. Then we have

$$\begin{aligned} ARE(SignedRank|t) &= \frac{e(T_{3n})}{e(T_{1n})} \\ &= 12\sigma^2\{\int_{-\infty}^{\infty} f^2(x)dx\}^2 \end{aligned}$$

As before to learn the performance of the competing tests, we compute ARE for different underlying density f. All these are provided in the table below. As earlier, we observe that the performances of Wilcoxon Signed rank test and the t statistic based test are more or less equivalent for normal, logistic and uniform parents. The improved efficiency is expected as

Table 2: **Comparing Signed rank Test and test based on t statistic**

$f(x)$	$ARE(SignedRank\|t)$
N(0,1)	$\frac{3}{\pi} = .96$
Logistic(0,1)	$\frac{\pi^2}{9} = 1.1$
Double Exponential(0,1)	1.5
R(-1,1)	1.0
Cauchy(0,1)	∞

the Wilcoxon Signed rank statistic utilizes both the sign and magnitude of the observations. It is further observed that for most of the distributions, Wilcoxon signed rank test is almost same or more efficient than the t statistic based test. Again t statistic based test is no longer desired for double exponential distributions and hence for such parents the efficiency of the Wilcoxon Signed rank test is significantly higher than its competitor. However, in a seminal work, Hodges and Lehmann(1956) proved that for every continuous density $f(x)$, $ARE(SignedRank|t)$ is at least .864.

Lastly, we consider comparing two completely nonparametric tests, namely, the Sign test and the Wilcoxon Signed rank test. Naturally , we arrive at the following expression of ARE

$$\begin{aligned} ARE(Sign|SignedRank) &= \frac{e(T_{2n})}{e(T_{3n})} \\ &= \frac{f^2(0)}{3\{\int_{-\infty}^{\infty} f^2(x)dx\}^2} \end{aligned}$$

It is interesting to observe that note that

$$\begin{aligned} ARE(Sign|SignedRank) &= ARE(Sign|t)ARE(t|SignedRank) \\ &= \frac{ARE(Sign|t)}{ARE(SignedRank|t)} \end{aligned}$$

Thus we immediately get the following table for different choices of f. The results are in agreement with our anticipation. Except for the double exponential distribution, the Sign

Table 3: **Comparing Sign Test and Wilcoxon Signed Rank Test**

$f(x)$	$ARE(Sign\|SignedRank)$
N(0,1)	$\frac{2}{3} = .67$
Logistic(0,1)	$\frac{3}{4} = .75$
Double Exponential(0,1)	1.33
R(-1,1)	.33

test is less efficient than the Wilcoxon Signed rank test. We note that Sign test statistic only uses the sign of the differences whereas Wilcoxon Signed rank statistic utilizes both the sign and magnitude of the observations. The use of excess information makes the latter more efficient.

3 Finally..

We have thus derived the theoretical expression of asymptotic power and consequently asymptotic relative efficiency. Although we restrict ourselves to the univariate distributions b ut ARE can also be computed for paired and two sample cases. Along the same line of thinking, the ARE between Mann-Whitney test and the test based on t statistic can be computed. However, we left these as an easy exercise for the reader. In addition, we have discussed so far only Pitman's notion of ARE but there are alternative developments by Hodges and Lehmann(1956) and Bahadur(1960). We, therefore, suggest the interested reader to go through the book by Gibbons and Chakraborti(2004) and the references therein to get more exposure in this context.

Large sample Inference: Module 7[1]

What we provide in this module

- Uniform convergence
- Examples
- A mathematical definition of uniform convergence with examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Uniform convergence

We have already discussed pointwise convergence of sequence of functions. It is seen that for pointwise convergence, it is necessary to get a feasible $\nu(x, \epsilon)$. So the question is natural, whether we can find a $\nu' = \nu'(\epsilon)$ depending only on ϵ such that the convergence holds. This leads to the concept of uniform convergence.

We explain the concept through an example. Consider $f_n(x) = x^n/n^2, x \in (-1, 1)$. Then it is already shown that $f_n(x) \to 0$ pointwise on (-1,1). One can recall that, for the above example we get a choice $\nu(x, \epsilon) = [1/\sqrt(\epsilon)] + 1$. Naturally such a choice is independent of x. Thus we get a common value of $\nu(x, \epsilon)$ ensuring pointwise convergence. Hence the convergence is more that pointwise and termed uniform(in the sense that a uniform choice of $\nu(x, \epsilon)$ exists).

1.1 Uniform convergence-Definition

Suppose $\{f_n(x)\}$ is a sequence of functions on A to R. Then $\{f_n(x)\}$ is said to converge uniformly on A if for any given $\epsilon > 0$, there exists an integer $\nu' = \nu'(\epsilon)$ such that for all $x \in A$ and $n \geq \nu'$,

$$|f_n(x) - f(x)| < \epsilon$$

is satisfied. One can use the symbol $f_n \rightrightarrows f$ to indicate uniform convergence on A. It is easy to observe that uniform convergence implies pointwise convergence and uniform convergence on A implies that on B where $B \subseteq A$.

Now we shall discuss such a concept graphically. Suppose $\{f_n(x)\}$ is a sequence of functions on A to R such that $f_n \rightrightarrows f$ on A. Then uniform convergence $f_n \rightrightarrows f$ can be thought of the existence of some ν', independent of x, such that the graph of f_n will lie inside the band $(f - \epsilon, f + \epsilon)$ for every $n \geq \nu'$.

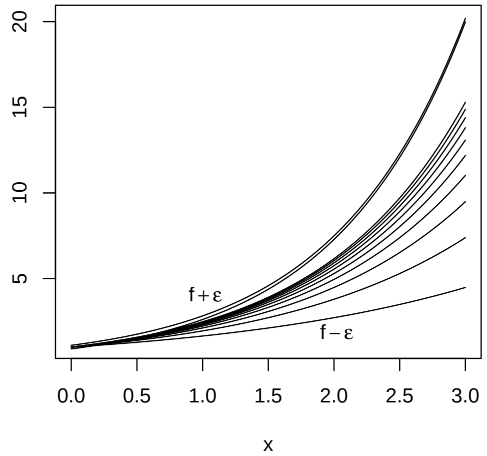


Figure 1: **Uniform Convergence**

1.2 Uniform Convergence: How to check in practice?

Suppose using some intuitive method, a limit function f is decided and pointwise convergence is established. Also suppose we get some $\nu(x, \epsilon)$ for $\epsilon > 0$ and $x \in A$. Consider $a, b \in A$, then for a given $\epsilon > 0$ we get real numbers $\nu(a, \epsilon)$ and $\nu(b, \epsilon)$ such that $|f_n(a) - f(a)| < \epsilon \ \forall \ n \geq \nu(a, \epsilon)$ and $|f_n(b) - f(b)| < \epsilon \ \forall \ n \geq \nu(b, \epsilon)$ are satisfied. Naturally, both the inequalities are satisfied if $n \geq max(\nu(a, \epsilon), \nu(b, \epsilon))$. The above suggests that if a feasible $\nu'(\epsilon) = \sup_{x \in B} \nu(x, \epsilon)$ exists for some $B \subseteq A$, then $f_n \rightrightarrows f$ on B.

1.3 Examples

Example 1: Consider $f_n(x) = x^2/n, x \in R$. Then for fixed x, we find that $\lim f_n(x) = 0$ and hence we claim $f(x) = 0 \,\forall\, x \in R$. In our previous module, we have determined $\nu(x, \epsilon) = [x^2/\epsilon] + 1$. Naturally $\sup_{x \in R} \nu(x, \epsilon)$ is not finite but for example $\sup_{x \in (0,a)} \nu(x, \epsilon) = [a^2/\epsilon] + 1$ exists for any finite $a > 0$. Thus $f_n \rightrightarrows 0$ on $(0, a)$ whereas $f_n \to 0$ pointwise for $x \in R$.

Example 2: Consider $f_n(x) = x^n, x \in (0, 1)$. Then using the methods previously

discussed we find that $\lim f_n(x) = f(x)$ pointwise on (0,1) for $f(x) = 0 \ \forall\ x \in (0,1)$.The following $\nu(x, \epsilon) = [log(\epsilon)/log(x)] + 1$ was derived. Now $\lim_{x \to 1-} \nu(x, \epsilon) = +\infty$ and hence convergence is not uniform on (0,1). However, $\nu(x, \epsilon)$ is increasing in x and hence if $0 < b < 1$, then $\sup_{x \in (0,b)} \nu(x, \epsilon) = \nu(b, \epsilon)$ is finite. Therefore, the convergence is uniform over $(0, b), b < 1$.

Example 3: Consider $f_n(x) = x^n/n^2, x \in (-1, 1)$. Then we have seen that pointwise convergence to $f(x) = 0 \ \forall\ x \in R$ is satisfied whenever $n \geq [1/\sqrt(\epsilon)] + 1$. Thus, in this case $\nu(x, \epsilon) = [1/\sqrt(\epsilon)] + 1$, which is independent of x. Hence the convergence is uniform over R. Thus we observe from the previous examples that pointwise convergence on A does not necessarily imply uniform convergence on A. But we can get a $B \subseteq A$ on which convergence is uniform.

Example 4: Consider $f_n(x) = nx/(x + n), x > 0$. Then it is already shown that $f_n(x) \to x \ \forall\ x > 0$ pointwise. The corresponding choice for $\nu(x, \epsilon)$ came out as $[\epsilon/x^2] + 1$. Since $\lim_{x \to 0+} \epsilon/x^2$ does not exist, convergence can not be uniform over $(0, \infty)$. However, $1/x^2$ is a decreasing function and hence $\sup_{x \in B} \nu(x, \epsilon)$ is finite for $B = [a, b)$ with $a > 0$. Hence the convergence is uniform over $[a, b)$ for some positive a.

Example 5: Consider $f_n(x) = nx/(x^2 + n^2), x > 0$. hen it is shown that $f_n \to f(x)$ for $f(x) = 0 \ \forall\ x > 0$ pointwise on $(0, \infty)$. In addition, we get the corresponding choice $\nu(x, \epsilon) = [\epsilon/x^2] + 1$. Then just as in the previous example, it can be shown that convergence is uniform over $[a, b)$ for some positive a. Similarly, uniform convergence for the other examples can also be investigated.

1.4 Uniform convergence-A mathematical definition

We have discussed so far uniform convergence from first principles. However, such a definition is not always easy to check and hence we need a simpler condition. Suppose $\{f_n(x)\}$ is a sequence of functions on A to R which converges pointwise to some $f(x)$ on A. Then

$f_n(x) \rightrightarrows f(x)$ on $B \subseteq A$ iff

$$M_n = \sup_{x \in B} |f_n(x) - f(x)| \to 0$$

as $n \to \infty$.

Consequently, we have the following results on sum and product functions.

The first result is on the uniform convergence of a sequence of functions, where the n th term is a sum of two. Formally if $f_n(x) \rightrightarrows f(x)$ and $g_n(x) \rightrightarrows g(x)$ uniformly on A then

(i) $cf_n(x) \rightrightarrows cf(x)$ on A and

(ii) $f_n(x) + g_n(x) \rightrightarrows f(x) + g(x)$ on A.

The next result is on the uniform convergence of a sequence of functions, where the n th term is a product of two. Formally if $f_n(x) \rightrightarrows f(x)$ and $g_n(x) \rightrightarrows g(x)$ uniformly on A and there exists $M > 0$ such that $|f_n(x)| \leq M$ and $|g_n(x)| \leq M$ then

$$f_n(x)g_n(x) \rightrightarrows f(x)g(x) \text{ on A.}$$

1.5 Examples

Example 1: Consider $f_n(x) = nxexp(-nx), x \geq 0$. We have already seen that $f_n(x) \to f(x)$ pointwise on $x \geq 0$, where $f(x) = 0, x \geq 0$. Now $M_n = \sup_{x \geq 0} nxexp(-nx)$ and a simple algebra shows that $nxexp(-nx)$ has a unique maximum at $x = 1/n$. Thus $M_n = exp(-1)$which is non zero and independent of n. Thus the convergence is not uniform over $[0, \infty)$. However, if we consider $B = [1, \infty]$, then $M_n \to 0$ and hence the convergence is uniform on B.

Example 2: Consider $f_n(x) = x^n, x \in [0, 1]$. Define $f(x) = 0$ if $0 \leq x < 1$ and $f(x) = 1$ for $x = 1$. Naturally the convergence holds trivially for x=0,1.However, we have already seen that $f_n(x) \to f(x)$ pointwise on $x \in [0, 1]$. Now $M_n = \sup_{x \in [0,1]} x^n$. Since x^n is an increasing function, it has a unique maximum at $x = 1$. Thus $M_n = 1$, which is different from zero and

hence the convergence is not uniform over $[0,1]$. However, if we consider $B=[0,a], a<1$, then $M_n \to 0$ and hence the convergence is uniform on B.

Example 3: Consider $f_n(x) = I(|x| \le n)$, where I is an indicator function. Then it is already shown that $\lim f_n(x) = 1 \ \forall x$ pointwise. Now $|f_n(x) - f(x)| = I(|x| > n)$. Now for every $n \ge 1$, $I(|x| > n)$ is either zero or 1. Thus $M_n = 1 \nrightarrow 0$ and hence the convergence is not uniform.

Example 4: Consider $f_n(x) = x^2 exp(-nx), x \in [0,1]$. It is easy to observe that $f_n(x) \to f(x)$ pointwise on $[0,1]$, where $f(x) = 0, x \in [0,1]$. Now $M_n = \sup_{x \in [0,1]} x^2 exp(-nx)$. A simple algebra shows that $x^2 exp(-nx)$ has a unique maximum at $x = 2/n$. Thus $M_n = \frac{4}{n^2} exp(-2) \to 0$ and hence the convergence is not uniform over $[0,1]$.

Large sample Inference: Module 2[1]

What we provide in this module

- Convergence concepts
- Examples
- Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Sequence & their Convergence

It is already discussed that large sample inference is attached with the performance evaluation when sample size becomes large. Naturally, large sample methods require the concepts of mathematical and stochastic convergence. Thus study of asymptotic properties in statistics depends heavily on results and concepts from real analysis and calculus. So we start with a in-depth review of the basic mathematical tools starting from sequence and series.

1.1 Sequence of real numbers

A sequence is a function from N, the set of positive integers to the real line R. Thus a sequence is a set of real numbers given by $a_n, n \geq 1$ such that $a_n : N \rightarrow R$. That is for every $n \in N$, $a_n \in R$. Naturally, the domain in this case (i.e. N) is countable and the range is the whole real line R. A sequence, therefore, generate sequentially for increasing values of $n \in N$.

As examples, one can thought of $a_n = 1/n^2$, $a_n = sin(n)$, $a_n = n$ and $a_n = (-1)^n$. Consider $a_n = 1/n^2$, then we have the values $a_1 = 1$,$a_2 = 1/4$ and so on. Thus we find that the value of a_n decreases with n. As another example, if we consider $a_n = n$, we find that the value increases with n. However for $a_n = (-1)^n$, we get that the sequence is either +1 or -1, i.e oscillates between -1 and +1. Thus we find that, sequences may be increasing or decreasing or bounded.

1.1.1 Monotonic sequences

A given sequence $a_n, n \geq 1$ is monotonically increasing if $a_{n+1} \geq a_n$ for every n. Similarly, a sequence is monotonically decreasing if $a_{n+1} \leq a_n$ for every n. In either case the sequence is called monotonic. It is easy to observe that n, $\sqrt{n}$, e^n are examples of increasing sequences. On the other hand $1/n^2$, $1/(\sqrt{n}+1)$ are the examples of decreasing sequences.

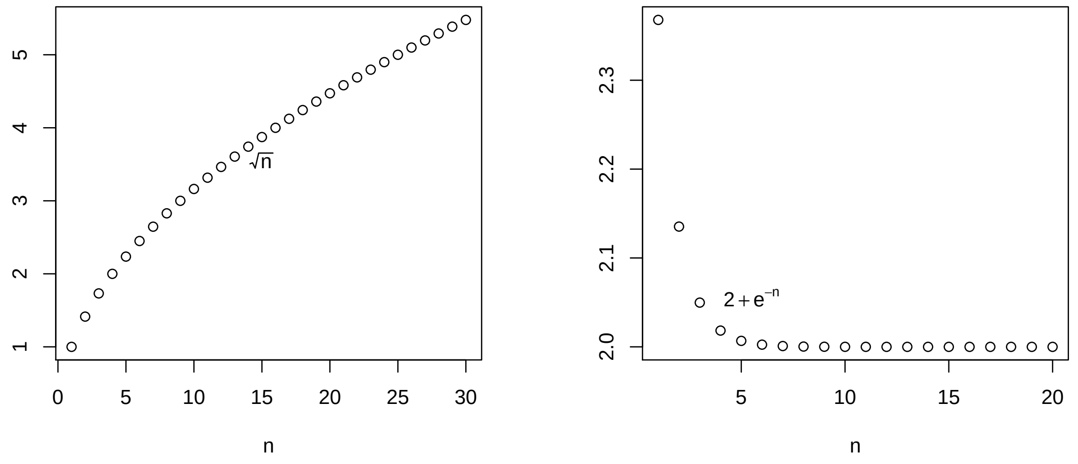


Figure 1: **Increasing & Decreasing Sequences**

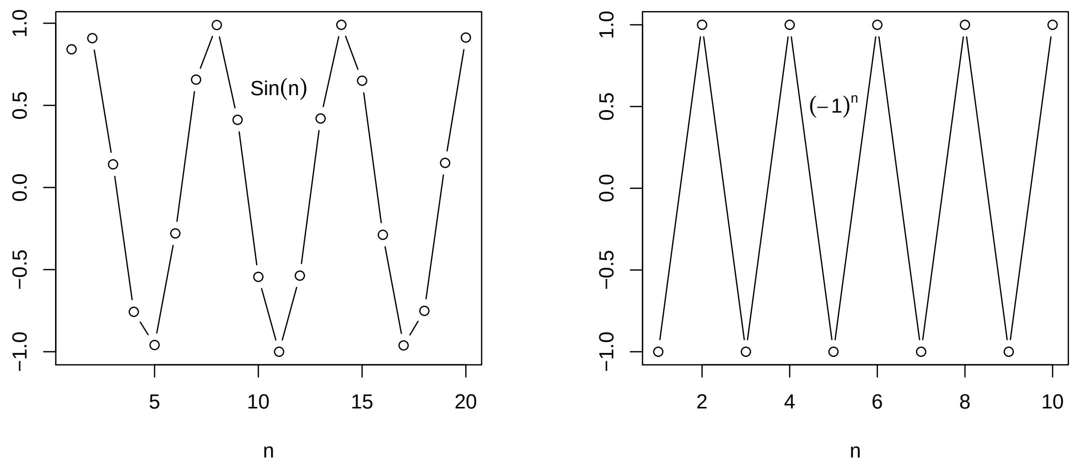


Figure 2: **Oscillating Sequences**

1.1.2 Bounded sequences

A sequence $a_n, n \geq 1$ is said to be bounded if there exists k_1 and k_2 such that

$$k_1 \leq a_n \leq k_2$$

is satisfied for every n, where k_1 and k_2 are independent of n. Since

$$k_1 \leq a_n \leq k_2 \Longrightarrow -|k_1| - |k_2| \leq a_n \leq |k_1| + |k_2|$$

we get that $|a_n| \leq k$ for some k. Thus bounded implies $|a_n| \leq k$ for some k. a given sequence. For example, we have $(-1)^n$ and $Sin(n)$.

1.1.3 Subsequence

Suppose $a_n, n \geq 1$ is a given sequence and $k(n)$ is an increasing sequence of positive integers. That is, $k(n+1) \geq k(n) \forall n$. Then the new sequence $\{a_{k(n)}\}$ is called a subsequence of $a_n, n \geq 1$. For example consider a sequence $a_n = 1/n$ and $k(n) = 2n$.Then the corresponding subsequence is $1/2n, n \geq 1$. Now consider a sequence $a_n = (-1)^n$ and $k(n) = 2n$. Then the corresponding subsequence is a constant sequence containing only 1.

1.2 Convergence of a sequence

Before going to the formal definition, let us consider some illustrative examples. Consider two sequences $1/n^2, n \geq 1$ and $Sin(n)/n, n \geq 1$. Plot will be helpful to study the behaviour of these sequences separately as we increase n.

It is easy to observe that, the curve decreases with increase in n. Finally, after a large value of n, the curve of $1/n^2$ becomes very close to the zero mark. However, for the second sequence, the curve wastes first few n to become stable. But as earlier, the curve maintains a little distance with the zero value for large n. Thus, in each case the value of the sequence becomes very close to a postulated value0, in these cases) as we exceed a large value of n. This gives the basis to define the convergence of a sequence.

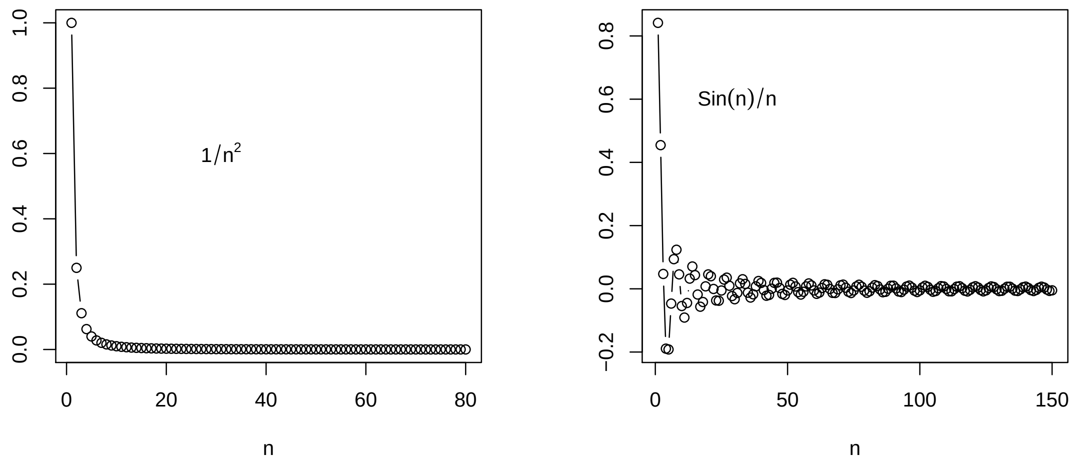


Figure 3: **Convergence**

1.2.1 Convergent sequence

A sequence $a_n, n \geq 1$ is said to be convergent if there exists $a \in R$ such that for every $\epsilon > 0$, there exists a positive integer $\nu = \nu(\epsilon)$ for which $|a_n - a| < \epsilon$ is satisfied for every $n \geq \nu$. If such an a exists, we write

$$\lim_{n \to \infty} a_n = a$$

and define a as the limit of the sequence. Let us consider few examples.

Example 1: Consider $a_n = 1 + exp(-n)$, which is a decreasing sequence. Then by inspection, we identify a as 1. Thus if a feasible $\nu(\epsilon)$ exists for every $\epsilon > 0$, then $\lim a_n = 1$. For $\epsilon \geq 1$, $|1 + exp(-n) - 1| < \epsilon$ is satisfied for any $n \geq 1$. However, if $0\epsilon < 1$, then $|1 + exp(-n) - 1| < \epsilon$ gives $n \geq [-log(\epsilon)] + 1$. Combining we set $\nu(\epsilon) = [-log(\epsilon)] + 1$. Since for every $\epsilon > 0$, we get a feasible choice, $\lim a_n = 1$.

Example 2: Consider $a_n = 1/n$, which is a decreasing sequence. Then by inspection, we identify a as 0. Thus if a feasible $\nu(\epsilon)$ exists for every $\epsilon > 0$, then $\lim a_n = 0$. If $\epsilon \geq 1$ then $|1/n - 0| \leq \epsilon$ for any $n \geq 1$. However, if $0 < \epsilon < 1$, then $|1/n - 0| < \epsilon$ gives $n \geq [1/\epsilon] + 1$.

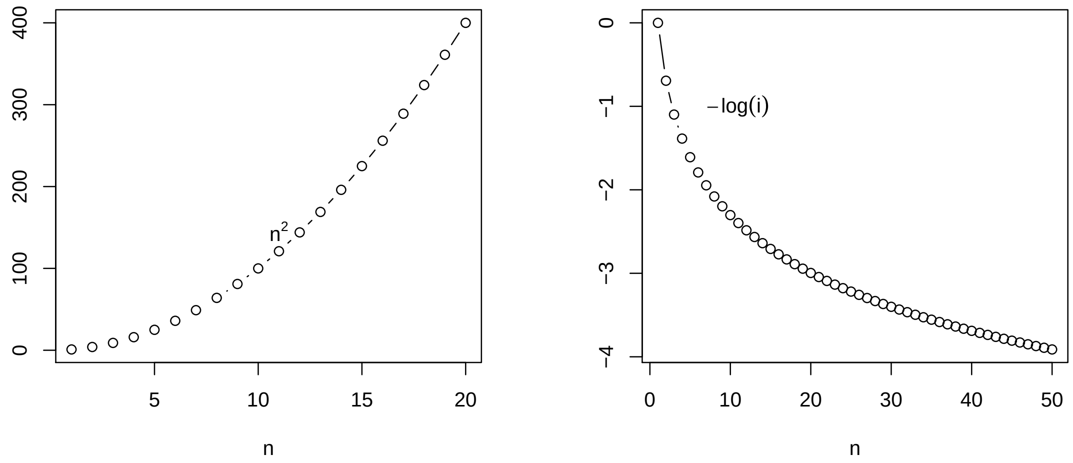


Figure 4: **Divergence**

Thus we set $\nu(\epsilon) = [1/\epsilon] + 1$. Since for every $\epsilon > 0$, we get a feasible choice, $\lim a_n = 0$.

Example 3: Consider $a_n = \sqrt{n+1} - \sqrt{n}$. Rationalizing we get $a_n = (\sqrt{n+1} + \sqrt{n})^{-1}$, which is a decreasing sequence. Then by inspection, we identify a as 0. Now $(\sqrt{n+1} + \sqrt{n})^{-1} < 1/\sqrt{n}$ and hence $|a_n - 0| < \epsilon$ will be satisfied if $|1/\sqrt{n} - 0| < \epsilon$ is satisfied. The last inequality gives $n \geq [1/\epsilon^2] + 1$ for $\epsilon < 1$. For $\epsilon \geq 1$, the inequality is satisfied for any $n \geq 1$. Thus we can set $\nu(\epsilon) = [1/\epsilon^2] + 1$. Since for every $\epsilon > 0$, we get a feasible choice, $\lim a_n = 0$.

1.2.2 Divergent sequence

From the analogy, if a sequence is not convergent, it is divergent. Let us consider some examples. Consider two sequences $n^2, n \geq 1$ and $-log(n), n \geq 1$. Plots will be helpful to study the behaviour of these sequences separately as we increase n.

It is easy to observe that, $a_n = n^2$ increases enormously whereas $a_n = -log(n)$ decreases as we increase n. Naturally, we can observe any finite number to which a_n converges. In

fact, after a large value of n, a_n becomes infinitely large or small. This gives the idea of the converse of the convergence, that is, divergence.

Thus formally, a sequence $a_n, n \geq 1$ diverges to $+\infty$ if for any $M(>0)$ there exists $\nu(M)$ such that

$$a_n > M \ \ \forall \ \ n \geq \nu(M)$$

is satisfied. On the other hand $a_n, n \geq 1$ diverges to $-\infty$ if $-a_n, n \geq 1$ diverges to $+\infty$. Let us illustrate the idea through an example. Consider $a_n = n^2$. Then $a_n > M$ implies $n \geq [\sqrt{M}] + 1$. Thus setting $\nu(M) = [\sqrt{M}] + 1$, we get that n^2 diverges to $+\infty$. Note that $(-1)^n$ is also an example of a convergent sequence as it does not converge to a fixed value but oscillates between -1 and +1.

2 Few results on convergence

Now we shall provide few convergence results without proof, which are useful in further development.

1. A convergent sequence is always bounded.
2. Limit of a convergent sequence is unique.
3. A sequence is convergent iff every subsequence from it is convergent.
4. A monotone sequence is convergent iff it is bounded.
5. If $a_n \geq b \ \forall \ n \geq 1$ for fixed b, $\lim a_n \geq b$ provided the limit exists.
6. If $\lim a_n = a$ and $\lim b_n = b$ then

   (i) $\alpha a_n + b_n \to a\alpha + b$

   (ii) $a_n b_n \to ab$

   (iii) $f(a_n) \to f(a)$ for continuous function f.

For a detailed proof, we refer the reader to the book by Rudin(1976). Next we shall give two useful approximations. The first one is Stirling's approximation, given by,

$$n! \approx \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}.$$

The next one is on a limit result on exponential approximation, given by,

$$\lim(1 + \frac{c}{n})^n = e^c$$

for fixed $c \in R$.

3 Applications in large sample theory

Now we shall discuss an application of the above results in large sample theory. Suppose X has a Poisson distribution with mean 5 and we are interested in P(X=n) for a large n. Observe that for $n > 5$, $5^n/n! < 5^5/5!(5/n)$. Thus $\lim 5^n/n! = 0$ and hence $\lim P(X = n) = 0$. Thus for large n the probability is negligibly small. Let us compute $P(X = n)$ for different values of n.

n	20	30	40
P(X=n)	2.64X 10^{-7}	2.36 X 10^{-14}	7.5 X 10^{-23}

From the above table it is easy to observe that the limiting value of zero is reached for n exceeding 30.

Large sample Inference: Module 3[1]

What we provide in this module

- Convergence of Series
- Applications in large sample theory
- Bounds of sequences

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Series & their Convergence

Series is a very widely applied concept in statistics. For example, Poisson , Geometric and Log series distributions are well known in statistical literature. It is interesting to observe that each of them is derived from an infinite series. For example, from the identity

$$-ln(1-\theta) = \sum_{x=0}^{\infty} \frac{\theta^x}{x!}$$

we get the pmf of log-series distribution

$$P(X=x) = -\frac{\theta^x}{x!ln(1-\theta)}, x = 0, 1, 2, ...$$

1.1 Series

For a detailed development, consider a sequence of real numbers $\{a_n\}$. Define another sequence as as $s_n = \sum_{k=1}^{n} a_k, n \geq 1$. The pair of sequences $\{a_n\}$ and $\{s_n\}$ define an infinite series. A series is defined by $\sum_{n=1}^{\infty} a_n$, where a_n is called the n th term of the series and s_n is the n th partial sum of the series.

1.1.1 Convergence of series

$\sum_{n=1}^{\infty} a_n$ is ordinarily convergent or divergent according as the sequence $\{s_n\}$ is convergent or divergent. Suppose $\{s_n\}$ is such that $\lim s_n = s$, where we allow the possibility of $s = \pm\infty$. If s is finite, we say that the series converges ordinarily and then $\sum_{n=1}^{\infty} a_n = s$. However, if $s = \pm\infty$, the series diverges and we say that the series diverges to $+\infty$ or $-\infty$.

We further have two other modes of convergence for a series containing both positive and negative terms. $\sum_{n=1}^{\infty} a_n$ is absolutely convergence if the series containing the absolute values of a_n is convergent. That is $\sum_{n=1}^{\infty} a_n$ is absolutely convergence if $\sum_{n=1}^{\infty} |a_n|$ is convergent. Again $\sum_{n=1}^{\infty} a_n$ is conditionally convergent if $\sum_{n=1}^{\infty} a_n$ is convergent but $\sum_{n=1}^{\infty} |a_n|$ is divergent. Note that Absolute convergence$\Rightarrow$ Ordinary convergence but the converse is not

always true.

Example 1: Consider the series $\sum_{n=1}^{\infty} n^{-2}$. Then the n th partial sum is simply

$$s_n = 1 + 2^{-2} + .. + n^{-2}.$$

Note that $\{s_n\}$ is an increasing sequence. Thus $2^{n+1} - 1 > n$ implies $s_n < s_{2^{n+1}-1}$. However, it can be shown that $s_{2^{n+1}-1} < 2 - 2^{-n} < 2 \ \forall \ n \geq 1$. Thus $s_n < 2 \ \forall \ n \geq 1$ and also $s_n \geq 1 \ \forall \ n \geq 1$. Hence $\{s_n\}$ is monotone and bounded. Thus by MCT, $\{s_n\}$ and hence the series converges.

Example 2: Consider the series $\sum_{n=1}^{\infty}(-1)^n$. Then the n th partial sum is either 0 for even n or -1 for odd n. Thus

$$s_n = \frac{-1 + (-1)^n}{2}.$$

Naturally $\{s_n\}$ is oscillating and hence does not converge. Then by definition, the series diverges. The series also diverges absolutely.

Example 3: Consider the geometric series $\sum_{n=1}^{\infty} r^{n-1}$. Then the n th partial sum s_n is n for $r = 1$ and $\frac{r^n-1}{r-1}$ for $r \neq 1$. If $|r| < 1$, $s_n \to (1-r)^{-1}$. However, if $r > 1$, $r^n \to \infty$ and hence $s_n \to \infty$. If $r \leq -1$, s_n oscillates. Thus combining we get that the geometric series converges for $|r| < 1$ and diverges otherwise.

Example 4: The p series $\sum_{n=1}^{\infty} n^{-p}$ is known to converge iff p¿1. Thus the harmonic series $\sum_{n=1}^{\infty} n^{-1}$ diverges.

Example 5: Consider the series $1 - \frac{1}{2} + \frac{1}{3} + ...$ It is well known that the above series converges to $ln(2)$. However, the corresponding series with absolute values is nothing but the harmonic series. which diverges. Thus we find that $1 - \frac{1}{2} + \frac{1}{3} + ..$ converges but the series with absolute values $\sum_{n=1}^{\infty} \frac{1}{n}$. Hence the series is conditionally convergent.

Example 6: Now we shall give an interesting example on rearranged series. In fact, the

terms of a conditionally convergent series may be suitably rearranged to converge or diverge. Consider the series $s = 1 - \frac{1}{2} + \frac{1}{3} + ..$, where $s = ln(2)$. The series is clearly conditionally convergent. Consider the rearranged series:

$$1 + (1/2 - 1) + 1/3 + (1/4 - 1/2) + 1/5 + (1/6 - 1/3) + ...$$

Now separating the positive and negative terms we get from the above

$$(1 + 1/2 + 1/3 + ..) - (1 + 1/2 + 1/3 + ..)$$

Now consider the following rearrangement of $\frac{s}{2}$:

$$1/2 - \frac{1}{4} + \frac{1}{6} + .. = 0 + 1/2 - 0 - 1/4 + 0 + 1/6 - 0 - 1/8 + ..$$

, Now adding with s, we get

$$s + \frac{s}{2} = (1 + 0) + (-1/2 + 1/2) + (0 + 1/3) + (-1/4 - 1/4) + (1/5 - 0) + ..$$

Thus the RHS is nothing but $1 + 1/3 - 1/2 + 1/5 + 1/7 - 1/4 + ...$ Then we get three possible values, corresponding to different rearrangements, of a conditionally convergent series as $ln(2)$, 0 and $3ln(2)/2$. For a detailed discussion on this issue, we refer the reader to *Methods of Real Analysis*(1963) by R.R. Goldberg.

2 Applications in probability theory

The above observation regarding rearranged series forms a basis to define the existence of moments.

Suppose X is random variable with the probability mass function $P(X = x_i) = p_i, i \geq 1$. Then for some function g, $Eg(X)$ is said to exist iff the series $\sum_i g(x_i)p_i$ is absolutely convergent. Now we shall give an example to show the usefulness of above requirement. Suppose $x_i = (-1)^{i-1}\frac{2^i}{i}, i \geq 1$ and $p_i = 2^{-i}$. Then the series $\sum_i x_i p_i$ is ordinarily convergent to $ln(2)$. However, the series is not convergent absolutely and hence E(X) does not exist though

$\sum_i x_i p_i$ is convergent.

Another application of convergence of series is to check the existence of moments of distributions with support as the whole set of integers. Suppose $P(X = n) = c\frac{1}{n(n+1)}, n = 1, 2, ..$ for some $c > 0$. Then c is such that $\sum_{n=1}^{\infty} P(X = n) = 1$. Now consider the n th partial sum of the above, which is $s_n = \sum_{i=1}^{n} c\frac{1}{i(i+1)}$. An adjustment using partial fractions, we get $s_n = c(1 - \frac{1}{n+1})$. Now $s_n \to c$ and hence, we get $\sum_{n=1}^{\infty} P(X = n) = 1$ only when $c = 1$. However, we find that the series $\sum_{n=1}^{\infty} nP(X = n)$ diverges and hence E(X) does not exist.

3 Supremum & Infimum of a sequence

We have already seen that the sequences can diverge. However, the asymptotic behavior of a divergent(mostly oscillatory) sequences can be described by considering the behavior of the bounds of the sequence. For $\{a_n\}$, l(u) is a lower(upper) bound if $a_n \geq l$ ($a_n \leq u$) $\forall\ n \geq 1$. From the properties of real numbers, if a sequence admits a lower(upper) bound, it also has a greatest(least) lower(upper) bound. The least upperbound and the greatest lowerbounds are defined, respectively, by $u_l = \sup_{n\geq 1} a_n$ and $l_u = \inf_{n\geq 1} a_n$. The limiting behavior of any sequence can be studied from the behaviour of these bounds for increasing n.

Note that $\sup_{k\geq n} a_n$ is a decreasing sequence and hence we get the limit as $\inf_{n\geq 1}\sup_{k\geq n} a_n$. The above quantity is known as limit superior of a_n and is denoted by either $\limsup a_n$ or $\overline{\lim} a_n$. Again $\inf_{k\geq n} a_n$ is an increasing sequence and we get the limit as $\sup_{n\geq 1}\inf_{k\geq n} a_n$. The above quantity is known as limit inferior of a_n and is denoted by either $\liminf a_n$ or $\underline{\lim} a_n$. Unlike limits, these two quantities always exist and can be used to judge the behaviour of a sequence in the limit.

Let us enumerate some results on these without proof.

1. $a_n, n \geq 1$ is convergent iff $\overline{\lim} a_n$ and $\underline{\lim} a_n$ are both finite and equal.

2. For any $a_n, n \geq 1$, $\overline{\lim} a_n \geq \underline{\lim} a_n$.

3. If $\overline{\lim} a_n$ is finite then $\overline{\lim} a_n = -\underline{lim}(-a_n)$

4. If $a_n, n \geq 1$ and $b_n, n \geq 1$ are two sequences with finite $\overline{\lim}$ and $\underline{\lim}$, then

$$(i)\ \overline{\lim}(a_n + b_n) \leq \overline{\lim} a_n + \overline{\lim} b_n$$

   and

$$(ii)\ \underline{\lim}(a_n + b_n) \geq \underline{\lim} a_n + \underline{\lim} b_n.$$

5. If $a_n, n \geq 1$ and $b_n, n \geq 1$ are two positive sequences with finite $\overline{\lim}$, then

$$\overline{\lim}(a_n b_n) \leq \overline{\lim} a_n \overline{\lim} b_n$$

   and

$$\underline{\lim}(a_n b_n) \geq \underline{\lim} a_n \underline{\lim} b_n$$

3.1 Examples

Example 1: Consider $a_n = e^{-n}, n \geq 1$. The sequence is strictly positive and monotonically decreasing in n. Now $\sup_{k \geq n} a_k = \sup_{k \geq n} e^{-k} = e^{-n}$. Then $\overline{\lim} a_n = \inf_{n \geq 1} e^{-n} = 0$. Again $\inf_{k \geq n} a_k = \inf_{k \geq n} e^{-k} = 0$. Thus $\underline{\lim} a_n = \sup_{n \geq 1} 0 = 0$. Thus $\overline{\lim} a_n = \underline{\lim} a_n = 0$ and hence the limit exists and $\lim a_n = 0$.

Example 2: Consider $a_n = 1 + (-1)^n, n \geq 1$. The sequence is non negative and oscillating between 0 and 1. Now $\sup_{k \geq n} a_k = 1 + \sup_{k \geq n} (-1)^k = 2$. Then $\overline{\lim} a_n = \inf_{n \geq 1} 2 = 2$. Again $\inf_{k \geq n} a_k = 1 + \inf_{k \geq n} (-1)^k = 0$. Thus $\underline{\lim} a_n = 0$ but $\overline{\lim} a_n = 2$. Since these two quantities are different, limit does not exist.

Large-Sample Inference: Module 4[1]

Learn More

1. Rohatgi, V.K. and Saleh, A.K.(2002). An introduction to probability and statistics. Second Edition, John Wiley & Sons Inc., New York.

2. Tom M. Apostol(1974). Mathematical Analysis. Addison-Wesley Publishing Company, Inc..

3. Robert G. Bartle and Donald R. Sherbert(1972). Introduction to Real Analysis, John Wiley & Sons Inc. New York.

4. Richard R. Goldberg(1964). Methods of Real Analysis, Blaisdell Publishing Company.

5. M.H. Protter and C.B. Morrey(1991). A First Course in Real Analysis, Springer-Verlag.

6. Kenneth A. Ross(1980). Elementary Real Analysis, Springer.

7. Walter Rudin(1976). Principles of Mathematical Analysis, Third Edition, McGraw Hill Inc.

8. Bandyopadhyay, S.(2011). Mathematical Analysis-Problems and Solutions, Academic Publishers, Kolkata.

9. Malik, S.C. and Arora, S.(2000). Mathematical Analysis, Second Edition, New Age International(P) Limited, Kolkata.

10. Mapa, S.K.(2004). Introduction to Real Analysis, Fourth Edition, Sarat Book Distributors, Kolkata.

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

Large sample Inference: Module 5[1]

What we provide in this module

- Further examples of series convergence
- Big O and Small o notations
- Gauss's test and examples
- Sequence of functions

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Further Examples

Example 1: Consider the series $\sum_{n=1}^{\infty}(n+1)^{-n}$. Then $a_n = (n+1)^{-n}$ and hence $(a_n)^{1/n} = (n+1)^{-1}$. Thus $\lim(a_n)^{1/n} = 0 < 1$ and hence by root test, the series converges.

Example 2: Consider the series $\sum_{n=1}^{\infty} n^n/n!$. Then $a_n = n^n/n!$ and hence $a_{n+1}/a_n = (1+1/n)^n$. It is well known that $\lim(1+1/n)^n = e > 1$. Hence by ratio test the series converges.

Example 3: Consider the series $1/3 + 3 + 1/3^2 + 3^2 + ...$ Then $a_{2n} = 3^n$ and $a_{2n+1} = 3^{-n}$, where $n = 0, 1, 2, ..$ and $\lim(a_{2n})^{1/2n} = \sqrt{3}$ and $\lim(a_{2n+1})^{1/2n+1} = 1/\sqrt{3}$. Hence $\overline{\lim} a_n = \sqrt{3} > 1$. Thus by root test the series is convergent.

Example 4: Consider the series $\sum_{n=2}^{\infty}(2 + log(n))^{-1}$. Naturally $a_n = 1/(2 + log(n)) = 1/(log(ne^2))$ and it is a decreasing quantity. Now $b^n a_{b^n} = b^n/(log(e^2 b^n)) = b^n/(2 + nlog(b))$. Since by definition $b > 1$, $b^n > 1$ and hence we get $b^n a_n = b^n/(2 + nlog(b)) > 1/(2 + nlog(b))$.Now b is a constant and $\sum_n 1/(2 + nlogb)$ is known to be divergent. Hence the original series diverges by Cauchy condensation test.

Example 5: Consider the series $\sum_{n=2}^{\infty}(nlog(n))^{-p}$. Naturally $a_n = 1/(nlog(n))^p$. Now $b^n a_{b^n} = b^n/(b^n log(b^n))^p = b^{n-np}/(nlog(b))^p$. Now for $p > 1$, $b^{n-np} < 1$. Since b is a constant, we get by comparison test that the series $\sum_n n^{-p}$ converges for $p > 1$. However for p=1, $b^n a_{b^n} = 1/(nlog(b))$, which is divergent. Now for $p < 1$, $b^{n-np} > 1$ and hence $b^n a_{b^n} > 1/(nlog(b))^p$. Since b is a constant, we get by comparison test that the series diverges for $p < 1$.

Example 6: [p series revisited] Consider the series $\sum_{n=1}^{\infty}(n)^{-p}$. For the p series, we observe $a_n = 1/n^p$ is decreasing in n. Now $b^n a_{b^n} = b^n/(b^n)^p = b^{n-np}$. Now for $p > 1$, $b^{1-p} < 1$ and hence $\sum b^n a_{b^n}$ becomes a geometric series with common ratio less than unity. Thus the series converges for $p > 1$. For p=1, $b^n a_{b^n} = 1$, which is a constant and therefore, by definition, the corresponding series diverges. Now for $p < 1$, $b^{1-p} > 1$ and hence from the properties of geometric series, the series diverges.

2 Growth orders

2.1 Big O notations

The symbols O was first introduced by E. Landau. The letter O is used to indicate growth rate of a function. Growth rate is commonly known as its order. These notations are often useful to compare growth rate and decrease rate of functions. Suppose a_n and b_n are two sequences, then $a_n = \bigcirc(b_n)$ as $n \to \infty$ iff $\frac{a_n}{b_n}$ is uniformly bounded for $n \geq n_0$. That is there exists a constant M and n_0 such that $|\frac{a_n}{b_n}| < M$ whenever $n \geq n_0$. However, $a_n = \bigcirc(1)$ implies that a_n is itself bounded for large n. If $a_n = \bigcirc(b_n) \implies \frac{a_n}{b_n} = \bigcirc(1)$, then a_n and b_n are said to grow or shrink in the same rate.

For example, consider the polynomial $a_0 + a_1 n + a_2 n^2 + a_3 n^3$. Then we can write the polynomial as $\bigcirc(n^3)$ whatever be the coefficients. However the polynomial $a_1 n^{-1} + a_2 n^{-2} + a_3 n^{-3}$ is represented by $\bigcirc(n^{-1})$. If a_n is a finite sum/difference of a number of functions, then the fastest growing quantity gives the order. Consider $a_n = n^2 + n + log(n)$, then naturally, the fastest growing quantity is n^2. Thus we can write $a_n = \bigcirc(n^2)$ as $n \to \infty$. On the other hand if a_n is a product of several factors, the constant term is omitted from the order relation.For example, consider the polynomial $a_n = 4 + 6n^3 - 2n^5$. Naturally the highest power is 5 and hence conventionally $a_n = \bigcirc(n^5)$. However, if we follow the definition, we need the auxiliary sequence b_n. Note that $|a_n| \leq 4 + 6n^3 + 2n^5 \leq 4 + 6n^5 + 2n^5 = 12n^5$ for $n \geq 1$. Thus we get M=12 , $n_0 = 1$ and $b_n = n^5$ On the other hand if a_n is a product of several factors, the constant term is omitted from the order relation.

Big O notation is useful in computer science to compare efficiency. Suppose a program of size n(in some sense) takes the time $a_n = n^2 + 2n$ to give the output. Another size n program for the same purpose takes the time $b_n = n^3 - n^2$ to give the output. Naturally the program taking less time is efficient. Now the latter program has the leading term n^3, which grows at a faster rate that the latter with increase in n. Thus big O notation can be use to indicate the time efficiency. Since $a_n = \bigcirc(n^2)$, it is more efficient.

2.2 Small o notations

The small o notation is used to capture the growth comparison of two quantities. If a_n and b_n are two sequences, then $a_n = \circ(b_n)$ as $n \to \infty$ if for every positive constant ϵ there exists $n_0 = n_0(\epsilon)$ such that $|\frac{a_n}{b_n}| < \epsilon$ for every $n \geq n_0$. This simply means that $a_n = \circ(b_n)$ iff $\lim \frac{a_n}{b_n} = 0$. Naturally small o is different from big O in the sense that , for the former, the inequality has to be true for some M whereas for the latter the inequality is to be satisfied for every positive ϵ. Thus $a_n = \circ(b_n) \Longrightarrow a_n = \bigcirc(b_n)$ but the converse is not true.

2.3 Tests related to order of growth: Gauss's test

Suppose $\sum_n a_n$ is a series with positive terms such that

$$\frac{a_n}{a_{n+1}} = 1 + \frac{\alpha}{n} + \frac{\beta_n}{n^p},$$

where $p > 1$ and β_n is a bounded sequence. Then the series converges if $\alpha > 1$ and diverges if $\alpha \leq 1$. Alternatively suppose $\sum_n a_n$ is a series with positive terms satisfying

$$\frac{a_n}{a_{n+1}} = 1 + \frac{\alpha}{n} + \bigcirc(n^{-p})$$

with $p > 1$. Then the series converges if $\alpha > 1$ and diverges if $\alpha \leq 1$.

As an example, consider the series $\sum_{n\geq 2} \frac{2^2 4^2 \ldots (2n-2)^2}{3^2 5^2 \ldots (2n-1)^2}$. Then it is easy to observe that $\lim \frac{a_{n+1}}{a_n} = 1$. Thus Ratio test fails. Now $\frac{a_n}{a_{n+1}} = 1 + 1/n + 1/4n^2 = 1 + 1/n + \bigcirc(n^{-2})$. Hence by Gauss test the series diverges.

As another example consider the p series $\sum_{n\geq 1} n^{-p}$. Then we get $a_n/a_{n+1} = (1 + 1/n)^p = 1 + p/n + \bigcirc(1/n^2)$. Thus the series converges for $p > 1$ and diverges for $p \leq 1$.

3 Sequence of functions

We have already discussed convergence of sequence and series. However, in many applications, we actually get a sequence or series involving some real variable x. For example, one

can consider, power series distributions, log series and e series as a function of x. Thus in addition to n there is an additional variable x. Naturally, convergence concepts are to be reformulated taking into account the presence of x.

3.1 Definition & convergence

Let A be a nonempty subset of R. Suppose for each $n \in N$, $f_n : A \to R$ be a function. Then $\{f_n(x)\}$ is defined as a sequence of functions on A to R. A is defined as the domain of the sequence of functions. $x^n, x \in [-1, 1]$, $exp(-nx), x \in R$ and $Sin(nx)/n, x \in R$ are examples of sequence of functions. It is easy to observe that for each fixed $x_0 \in A$, we get a real sequence $\{f_n(x_0)\}$.

We have already introduces the convergence concepts for a real sequence $\{a_n\}$. But in this case, we have an extra variable x. Thus it is intuitively clear that we need to develop separate convergence concepts depending on the role of x. For example, consider $x^n, x \in [-1, 1]$. For $x = 1/2$, we get the sequence $(1/2)^n$, which converges to zero for large n. However, if we take $x = -1$, the sequence oscillates between -1 and 1 and hence diverges. Thus depending on x, the sequence either converges or diverges.

Large sample Inference: Module 6[1]

What we provide in this module

- Pointwise convergence
- Examples & Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Pointwise convergence

We have already defined sequence of functions and the requirement of convergence. Now we provide the convergence concepts of a sequence of functions in details.

1.1 Definition

Suppose $\{f_n(x)\}$ is a sequence of functions on A to R. Then $\{f_n(x)\}$ is said to converge pointwise on A if for each $x \in A$, the sequence converges. Thus for every fixed $x_0 \in A$, the sequence $\{f_n(x_0)\}$ converges and hence there exists some $f(x_0)$ such that $\lim f_n(x_0) = f(x_0)$. Since x_0 is fixed, $f(x)$ exists for every $x \in A$. Such f is defined as the limit function of $\{f_n(x)\}$ and we write $f_n \to f$ pointwise.

1.2 How to get f?

Although we define an f theoretically but in practice getting such an f is not easy. We shall explain deciding an f using examples. Suppose $\{f_n(x)\} = x^2/n, x \in [0, 1]$ is a sequence of functions on A to R. Let us plot the functions for different values of n on the same graph paper. We, in addition, plot $\{f_n(x)\} = x^n, x \in [0, 1)$ for different values of n.

Thus we observe that for the first figure, the functions approach zero as we increase n. Then it is easy to take the limit function $f(x) = 0 \ \forall \ x \in [0, 1]$. Intuitively, the above is justified as for fixed x, $x^2/n \to 0$. However, for the second function we get a sequence of decreasing curves. Then it is easy to expect the horizontal axis as the limit. This suggests to take the limit function $f(x) = 0 \forall \ x \in [0, 1)$. Thus we find that after a large n, the graph becomes very close to the limit function. Obviously, such an n depends on the value of x. Thus $f_n \to f$ implies that for every $x \in A$ and for every positive ϵ, there exists a positive integer $\nu(x, \epsilon)$ such that $|f_n(x) - f(x)| < \epsilon$ is satisfied for every $n \geq \nu(x, \epsilon)$. However, looking

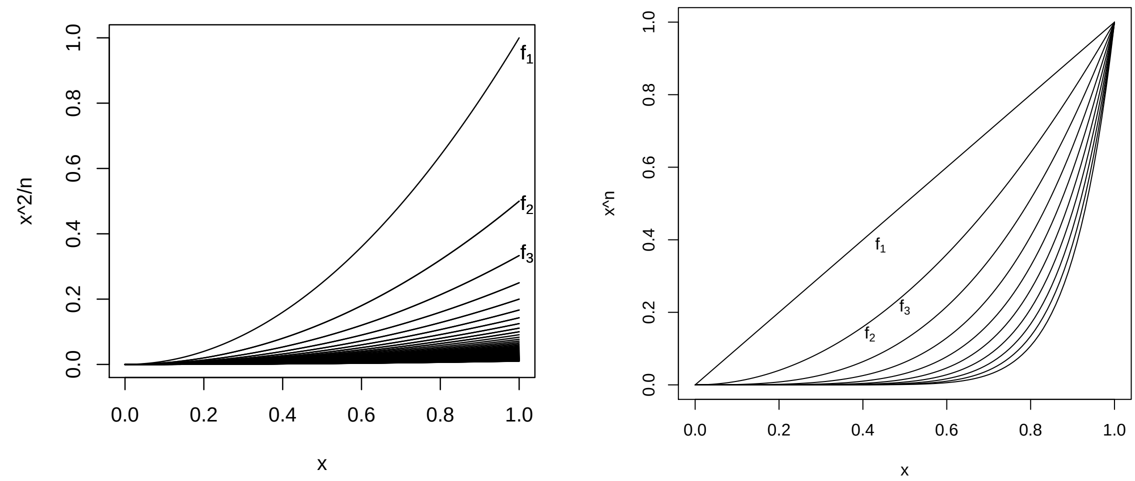


Figure 1: **Convergence for a sequence of functions**

at the nature of $\lim f_n(x)$ for fixed x, f is defined in practice.

1.3 Examples

Example 1: Consider $f_n(x) = x^2/n, x \in R$. Then for fixed x, we find that $\lim f_n(x) = 0$ and hence we claim $f(x) = 0 \; \forall \; x \in R$. However, the claim will be justified, if we can find a feasible $\nu(x, \epsilon)$, as defined above. Now $|f_n(x) - f(x)| < \epsilon$ implies $n \geq [x^2/\epsilon] + 1$ Thus we get $\nu(x, \epsilon) = [x^2/\epsilon] + 1$. Hence $\lim f_n(x) = f(x)$ pointwise for $f(x) = 0 \; \forall x \in R$.

Example 2: Consider $f_n(x) = x^n, x \in (0, 1)$. Then for fixed x, we find that x^n is a positive fraction and hence $\lim f_n(x) = 0$ and we claim $f(x) = 0 \; \forall \; x \in R$. For $\epsilon \geq 1$, $|f_n(x) - f(x)| < \epsilon$ is satisfied for any $n \geq 1$. However, for $\epsilon \in (0, 1)$, $|f_n(x) - f(x)| < \epsilon$ gives $n \geq [log(\epsilon)/log(x)] + 1$. Thus, in any case we can take $\nu(x, \epsilon) = [log(\epsilon)/log(x)] + 1$. Hence $\lim f_n(x) = f(x)$ pointwise.

Example 3: Consider $f_n(x) = x^n/n^2, x \in (-1,1)$. Then for fixed x, we find that $x^n \to 0$ so that $\lim f_n(x) = 0$ and we claim $f(x) = 0\ \forall\ x \in R$. However, finding a $\nu(x,\epsilon)$ from $|f_n(x) - f(x)| < \epsilon$ is quite complicated. Now it is easy to observe that $|x^n/n^2 - 0| < |1/n^2 - 0|$ for every $x \in (-1,1)$. Thus, $|1/n^2 - 0| < \epsilon \Rightarrow |x^n/n^2 - 0| < \epsilon$. Now for $\epsilon \geq 1$, $|1/n^2 - 0| < \epsilon$ is satisfied for every $n \geq 1$. However, for $\epsilon \in (0,1)$, we get $n \geq [1/\sqrt(\epsilon)] + 1$. Thus, in any case we can take $\nu(x,\epsilon) = [1/\sqrt(\epsilon)] + 1$. Hence $\lim f_n(x) = f(x)$ pointwise.

Example 4: Consider $f_n(x) = nx/(x+n), x \geq 0$. Then for fixed x, we find that $n/(x+n) \to 1$ so that $\lim f_n(x) = x$ and we claim $f(x) = x\ \forall\ x \geq 0$. Now $|f_n(x) - f(x)| = \frac{x^2}{x+n} < \frac{x^2}{n}$. Thus $|f_n(x) - f(x)| < \epsilon$ is satisfied, if $\frac{x^2}{n} < \epsilon$ is satisfied. Now $\frac{x^2}{n} < \epsilon$ gives $n \geq [\epsilon/x^2] + 1$. Thus, we define $\nu(x,\epsilon) = [\epsilon/x^2] + 1$, which is a feasible choice and hence $\lim f_n(x) = x, x \geq 0$ pointwise.

Example 5: Consider $f_n(x) = nx/(x^2+n^2), x \geq 0$. Then for fixed x, we find that $n/(x^2+n^2) \to 0$ so that $\lim f_n(x) = 0$ and we claim $f(x) = 0\ \forall\ x \geq 0$. Now $|f_n(x) - f(x)| = nx/(x^2+n^2) < \frac{x}{n}$. Thus $|f_n(x) - f(x)| < \epsilon$ is satisfied, if $\frac{x}{n} < \epsilon$ is satisfied. Now $\frac{x}{n} < \epsilon$ gives $n \geq [\epsilon/x] + 1$. Thus, if we define $\nu(x,\epsilon) = [\epsilon/x] + 1$, we get a feasible choice and hence $\lim f_n(x) = 0, x \geq 0$ pointwise.

Example 6: Consider $f_n(x) = nxexp(-nx), x \geq 0$. However, deciding an appropriate limit function is not an easy task. Thus, we plot the function for different values of n on the same graph paper to get an idea. The plot is given below.

It is easy to observe that the curve attains a peak and then diminishes to zero. As we increase n, the point at which peak is attained becomes nearer to zero. This suggests to take the limit function as $f(x) = 0, x \geq 0$. However finding a $\nu(x,\epsilon)$ from the relation $|f_n(x) - 0| < \epsilon$ is not an easy task. Again, $exp(nx) = 1 + nx + (nx)^2/2 + ...$ gives

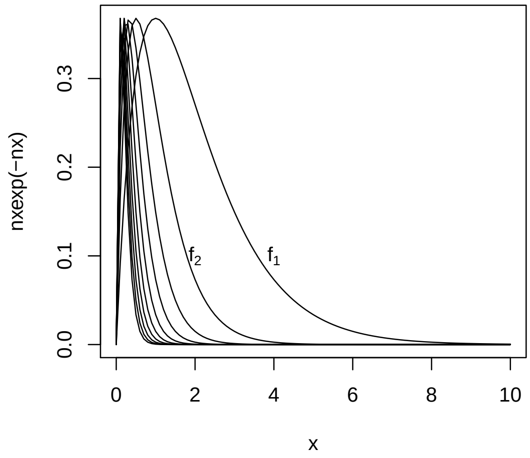


Figure 2: **Deciding limit function**

$exp(nx) > (nx)^2/2$. Thus $|f_n(x) - 0| < \epsilon$ will be satisfied if we can find a $\nu(x, \epsilon)$ satisfying $2/(nx) < \epsilon$. Now the latter inequality gives $\nu(x, \epsilon) = [2/(x\epsilon)] + 1$. Since we get a feasible choice, $\lim f_n(x) = 0, x \geq 0$ pointwise.

Example 7: Consider $f_n(x) = I(|x| \leq n)$, where I is an indicator function. As earlier, the first task is to decide a limit function f. Intuitively, it is clear that as we increase n, the more of the real line will be covered and we get the value unity there. Thus we claim $f(x) = 1 \ \forall \ x$. Now $|f_n(x) - f(x)| = I(|x| > n)$. Clearly for $\epsilon \geq 1$, $|f_n(x) - f(x)| < \epsilon$ is satisfied for every $n \geq 1$. For $\epsilon \in (0, 1)$, the above inequality is satisfied whenever $|x| \leq n$ or equivalently $n \geq \nu(x, \epsilon) = [|x|] + 1$. Hence $\lim f_n(x) = 1 \ \forall x$ pointwise.

Large sample Inference: Module 7[1]

What we provide in this module

- Uniform convergence
- Examples
- A mathematical definition of uniform convergence with examples

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Uniform convergence

We have already discussed pointwise convergence of sequence of functions. It is seen that for pointwise convergence, it is necessary to get a feasible $\nu(x, \epsilon)$. So the question is natural, whether we can find a $\nu' = \nu'(\epsilon)$ depending only on ϵ such that the convergence holds. This leads to the concept of uniform convergence.

We explain the concept through an example. Consider $f_n(x) = x^n/n^2, x \in (-1, 1)$. Then it is already shown that $f_n(x) \to 0$ pointwise on (-1,1). One can recall that, for the above example we get a choice $\nu(x, \epsilon) = [1/\sqrt{(\epsilon)}] + 1$. Naturally such a choice is independent of x. Thus we get a common value of $\nu(x, \epsilon)$ ensuring pointwise convergence. Hence the convergence is more that pointwise and termed uniform(in the sense that a uniform choice of $\nu(x, \epsilon)$ exists).

1.1 Uniform convergence-Definition

Suppose $\{f_n(x)\}$ is a sequence of functions on A to R. Then $\{f_n(x)\}$ is said to converge uniformly on A if for any given $\epsilon > 0$, there exists an integer $\nu' = \nu'(\epsilon)$ such that for all $x \in A$ and $n \geq \nu'$,

$$|f_n(x) - f(x)| < \epsilon$$

is satisfied. One can use the symbol $f_n \rightrightarrows f$ to indicate uniform convergence on A. It is easy to observe that uniform convergence implies pointwise convergence and uniform convergence on A implies that on B where $B \subseteq A$.

Now we shall discuss such a concept graphically. Suppose $\{f_n(x)\}$ is a sequence of functions on A to R such that $f_n \rightrightarrows f$ on A. Then uniform convergence $f_n \rightrightarrows f$ can be thought of the existence of some ν', independent of x, such that the graph of f_n will lie inside the band $(f - \epsilon, f + \epsilon)$ for every $n \geq \nu'$.

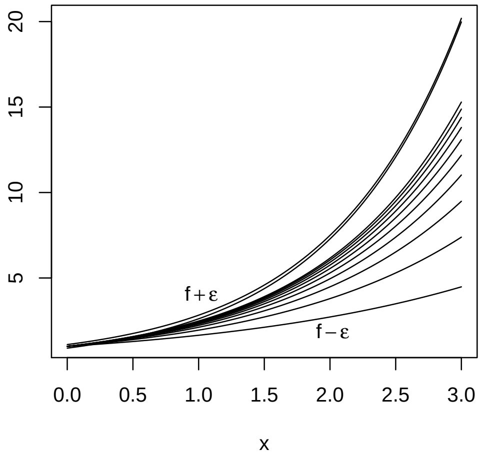


Figure 1: **Uniform Convergence**

1.2 Uniform Convergence: How to check in practice?

Suppose using some intuitive method, a limit function f is decided and pointwise convergence is established. Also suppose we get some $\nu(x, \epsilon)$ for $\epsilon > 0$ and $x \in A$. Consider $a, b \in A$, then for a given $\epsilon > 0$ we get real numbers $\nu(a, \epsilon)$ and $\nu(b, \epsilon)$ such that $|f_n(a) - f(a)| < \epsilon \; \forall \; n \geq \nu(a, \epsilon)$ and $|f_n(b) - f(b)| < \epsilon \; \forall \; n \geq \nu(b, \epsilon)$ are satisfied. Naturally, both the inequalities are satisfied if $n \geq max(\nu(a, \epsilon), \nu(b, \epsilon))$. The above suggests that if a feasible $\nu'(\epsilon) = \sup_{x \in B} \nu(x, \epsilon)$ exists for some $B \subseteq A$, then $f_n \rightrightarrows f$ on B.

1.3 Examples

Example 1: Consider $f_n(x) = x^2/n, x \in R$. Then for fixed x, we find that $\lim f_n(x) = 0$ and hence we claim $f(x) = 0 \; \forall \; x \in R$. In our previous module, we have determined $\nu(x, \epsilon) = [x^2/\epsilon] + 1$. Naturally $\sup_{x \in R} \nu(x, \epsilon)$ is not finite but for example $\sup_{x \in (0,a)} \nu(x, \epsilon) = [a^2/\epsilon] + 1$ exists for any finite $a > 0$. Thus $f_n \rightrightarrows 0$ on $(0, a)$ whereas $f_n \to 0$ pointwise for $x \in R$.

Example 2: Consider $f_n(x) = x^n, x \in (0, 1)$. Then using the methods previously

discussed we find that $\lim f_n(x) = f(x)$ pointwise on (0,1) for $f(x) = 0 \; \forall \; x \in (0,1)$.The following $\nu(x,\epsilon) = [log(\epsilon)/log(x)] + 1$ was derived. Now $\lim_{x \to 1-} \nu(x,\epsilon) = +\infty$ and hence convergence is not uniform on (0,1). However, $\nu(x,\epsilon)$ is increasing in x and hence if $0 < b < 1$, then $\sup_{x \in (0,b)} \nu(x,\epsilon) = \nu(b,\epsilon)$ is finite. Therefore, the convergence is uniform over $(0,b), b < 1$.

Example 3: Consider $f_n(x) = x^n/n^2, x \in (-1,1)$. Then we have seen that pointwise convergence to $f(x) = 0 \; \forall \; x \in R$ is satisfied whenever $n \geq [1/\sqrt{(\epsilon)}] + 1$. Thus, in this case $\nu(x,\epsilon) = [1/\sqrt{(\epsilon)}] + 1$, which is independent of x. Hence the convergence is uniform over R. Thus we observe from the previous examples that pointwise convergence on A does not necessarily imply uniform convergence on A. But we can get a $B \subseteq A$ on which convergence is uniform.

Example 4: Consider $f_n(x) = nx/(x+n), x > 0$. Then it is already shown that $f_n(x) \to x \; \forall \; x > 0$ pointwise. The corresponding choice for $\nu(x,\epsilon)$ came out as $[\epsilon/x^2] + 1$. Since $\lim_{x \to 0+} \epsilon/x^2$ does not exist, convergence can not be uniform over $(0,\infty)$. However, $1/x^2$ is a decreasing function and hence $\sup_{x \in B} \nu(x,\epsilon)$ is finite for $B = [a,b)$ with $a > 0$. Hence the convergence is uniform over $[a,b)$ for some positive a.

Example 5: Consider $f_n(x) = nx/(x^2+n^2), x > 0$. hen it is shown that $f_n \to f(x)$ for $f(x) = 0 \; \forall \; x > 0$ pointwise on $(0,\infty)$. In addition, we get the corresponding choice $\nu(x,\epsilon) = [\epsilon/x^2] + 1$. Then just as in the previous example, it can be shown that convergence is uniform over $[a,b)$ for some positive a. Similarly, uniform convergence for the other examples can also be investigated.

1.4 Uniform convergence-A mathematical definition

We have discussed so far uniform convergence from first principles. However, such a definition is not always easy to check and hence we need a simpler condition. Suppose $\{f_n(x)\}$ is a sequence of functions on A to R which converges pointwise to some $f(x)$ on A. Then

$f_n(x) \rightrightarrows f(x)$ on $B \subseteq A$ iff

$$M_n = \sup_{x \in B} |f_n(x) - f(x)| \to 0$$

as $n \to \infty$.

Consequently, we have the following results on sum and product functions.

The first result is on the uniform convergence of a sequence of functions, where the n th term is a sum of two. Formally if $f_n(x) \rightrightarrows f(x)$ and $g_n(x) \rightrightarrows g(x)$ uniformly on A then

(i) $cf_n(x) \rightrightarrows cf(x)$ on A and

(ii) $f_n(x) + g_n(x) \rightrightarrows f(x) + g(x)$ on A.

The next result is on the uniform convergence of a sequence of functions, where the n th term is a product of two. Formally if $f_n(x) \rightrightarrows f(x)$ and $g_n(x) \rightrightarrows g(x)$ uniformly on A and there exists $M > 0$ such that $|f_n(x)| \leq M$ and $|g_n(x)| \leq M$ then

$$f_n(x)g_n(x) \rightrightarrows f(x)g(x) \text{ on A.}$$

1.5 Examples

Example 1: Consider $f_n(x) = nxexp(-nx), x \geq 0$. We have already seen that $f_n(x) \to f(x)$ pointwise on $x \geq 0$, where $f(x) = 0, x \geq 0$. Now $M_n = \sup_{x \geq 0} nxexp(-nx)$ and a simple algebra shows that $nxexp(-nx)$ has a unique maximum at $x = 1/n$. Thus $M_n = exp(-1)$which is non zero and independent of n. Thus the convergence is not uniform over $[0, \infty)$. However, if we consider $B = [1, \infty]$, then $M_n \to 0$ and hence the convergence is uniform on B.

Example 2: Consider $f_n(x) = x^n, x \in [0, 1]$. Define $f(x) = 0$ if $0 \leq x < 1$ and $f(x) = 1$ for $x = 1$. Naturally the convergence holds trivially for x=0,1.However, we have already seen that $f_n(x) \to f(x)$ pointwise on $x \in [0, 1]$. Now $M_n = \sup_{x \in [0,1]} x^n$. Since x^n is an increasing function, it has a unique maximum at $x = 1$. Thus $M_n = 1$, which is different from zero and

hence the convergence is not uniform over $[0,1]$. However, if we consider $B=[0,a], a<1$, then $M_n \to 0$ and hence the convergence is uniform on B.

Example 3: Consider $f_n(x) = I(|x| \leq n)$, where I is an indicator function. Then it is already shown that $\lim f_n(x) = 1 \; \forall x$ pointwise. Now $|f_n(x) - f(x)| = I(|x| > n)$. Now for every $n \geq 1$, $I(|x| > n)$ is either zero or 1. Thus $M_n = 1 \nrightarrow 0$ and hence the convergence is not uniform.

Example 4: Consider $f_n(x) = x^2 exp(-nx), x \in [0,1]$. It is easy to observe that $f_n(x) \to f(x)$ pointwise on $[0,1]$, where $f(x) = 0, x \in [0,1]$. Now $M_n = \sup_{x \in [0,1]} x^2 exp(-nx)$. A simple algebra shows that $x^2 exp(-nx)$ has a unique maximum at $x = 2/n$. Thus $M_n = \frac{4}{n^2} exp(-2) \to 0$ and hence the convergence is not uniform over $[0,1]$.

Large sample Inference: Module 8[1]

What we provide in this module

- Consequences of uniform convergence for sequence of functions
- Series of functions & their convergence
- Examples
- Consequences of uniform convergence for series of functions

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Consequences of uniform convergence

Now we shall discuss few consequences of uniform convergence. In fact, uniform convergence allows exchangeability of different operations like limits, differentiation, integration, etc.

1.1 Uniform convergence & limit interchanging

If $f_n(x) \rightrightarrows f(x)$ on A, where f_n is continuous on A for every n. Then f is also continuous on A. Thus for any $a \in A$, we have

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty}\lim_{x\to a} f_n(x).$$

The implication of the above is that if the limits are not interchangeable, then uniform convergence does not hold. For example, consider $f_n(x) = x^n, x \in [0, 1)$ then the pointwise limit is the function $f(x) = 0, x \in [0, 1)$. However $\lim_{n\to\infty} f_n(x) = 0$. Thus $\lim_{x\to 1} f(x) \neq \lim_{n\to\infty}\lim_{x\to a} f_n(x)$. Hence the convergence can not be uniform.

1.2 Uniform convergence & differentiation

If $f_n(x)$ be a sequence of functions on [a,b] such that
(i) for some $x_0 \in (a, b)$, $f_n(x_0) \to f(x_0)$ and
(ii) $f_n'(x)$ exists at all $x \in [a, b]$ and converges uniformly to some function g on [a,b].
Then $f_n(x) \rightrightarrows f(x)$ on [a,b] and $g(x) = f'(x)\ \forall\ x \in [a, b]$. That is

$$\lim_{n\to\infty} \frac{df_n(x)}{dx} = \frac{d}{dx}\lim_{n\to\infty} f_n(x)\ \forall\ x \in [a, b].$$

For better understanding, consider an example with $f_n(x) = \frac{x}{1+n^2x^2}, |x| \leq 1$. Define $f(x) = 0, x \in [-1, 1]$, then it can be shown that $M_n = \sup_{x\in[-1,1]} |f_n(x) - f(x)| = 1/2n \to 0$ as $n \to \infty$. Thus the convergence is uniform. However $f_n'(x) = \frac{1-n^2x^2}{(1+n^2x^2)^2}$ and hence for $x = 0$, $\lim_{n\to\infty} f_n'(x) = 1$. Thus $\lim_{n\to\infty} f_n'(x) \neq f'(x)$ for x=0.

1.3 Uniform convergence & integration

Suppose $f_n(x)$ is a sequence of Riemann integrable functions on [a,b] such that $f_n(x) \rightrightarrows f(x)$ on [a,b]. Then f is also Riemann integrable on [a,b] and

$$\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b \lim_{n\to\infty} f_n(x)dx.$$

For an example, consider $f_n(x) = nxexp(-nx^2), x \in [0,1]$. Define $f(x) = 0, x \in [0,1]$, then it can be shown that $M_n = \sup_{x\in[0,1]} |f_n(x) - f(x)| = \sqrt{n/2} \to \infty$ as $n \to \infty$. Thus the convergence is not uniform. Thus $\int_0^1 f_n(x)dx = (1 - exp(-n))/2 \to 1/2$ but $\int_0^1 f(x)dx = 0$. However if we consider the interval [1,3], then the convergence is uniform and hence $\int_1^3 f_n(x)dx \to 0 = \int_1^3 f(x)dx$.

2 Series of functions-Motivation

We have already discussed sequence of functions and their convergence. Like ordinary series, the idea of sequence of functions can also be extended to series of functions. For example, we have well known e series and log series. As earlier, we can have different concepts of convergence.

Let $\{f_n(x)\}$ be a sequence of functions on A to R. Then $f_1+f_2+f_3+..$ is said to be a series or infinite series and is often denoted by $\sum_{n=1}^{\infty} f_n(x)$. Define the sequence of partial sums of the infinite series by $s_n(x) = \sum_{k=1}^{n} f_k(x)$. If the sequence $\{s_n(x)\}$ converges pointwise to some $s(x)$ on A, the series $\sum_{n=1}^{\infty} f_n(x)$ is said to converge pointwise to $s(x)$ on A. However, if the convergence is uniform to some $s(x)$ on A, the series $\sum_{n=1}^{\infty} f_n(x)$ converges uniformly to $s(x)$ on A. f the series $\sum_{n=1}^{\infty} |f_n(x)|$ converges for each $x \in A$, $\sum_{n=1}^{\infty} f_n(x)$ is said to converge absolutely on A.

2.1 Examples

Example 1: Consider the series: $x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + ...$, for $x \in [0,1]$. Then $s_n(x) = \sum_{i=0}^{n-1} \frac{x^2}{(1+x^2)^i} = (1+x^2) - (1+x^2)^{-n+1}$ is the n th partial sum. Thus we find that

$$
\begin{aligned}
\lim s_n(x) &= 0 \text{ if } x = 0 \\
&= 1 + x^2 \text{ if } x \in (0,1]
\end{aligned}
$$

Thus $s_n(x)$ converges pointwise to $s(x) = (1+x^2)I(0 < x \leq 1)$ with I as an indicator function. Hence the series converges pointwise to $s(x)$ on [0,1]. However, s(x) is not continuous on [0,1] but each $s_n(x)$ is continuous on [0,1] and hence the convergence is not uniform on [0,1].

Example 2: Consider the series: $\sum_{n=1}^{\infty} x^n(1-x)^n$, for $x \in [0,1]$. Then $s_n(x) = \sum_{i=1}^{n} x^i(1-x)^i = x(1-x)\frac{1-[x(1-x)]^{n-1}}{1-x(1-x)}$ is the n th partial sum. Thus we find that

$$
\begin{aligned}
s(x) = \lim s_n(x) &= 0 \text{ if } x = 0, 1 \\
&= \frac{x(1-x)}{1-x(1-x)} \text{ if } x \in (0,1)
\end{aligned}
$$

Then $s_n(x)$ converges pointwise to $s(x)$ and hence the series converges pointwise to $s(x)$ on [0,1].

Example 3: Consider the series: $\sum_{n=1}^{\infty} Sin(nx)/n^p$, for $x \in R$. Let us fix x at a and consider the convergence of $\sum_n Sin(na)/n^p$. Now $|Sin(na)/n^p| \leq 1/n^p$ for all $n \geq 1$. Hence by comparison test $\sum_n |Sin(na)|/n^p$ converges for $p > 1$, that is the series converges absolutely. Since a is arbitrary, the series $\sum_{n=1}^{\infty} Sin(nx)/n^p$ is absolutely convergent on R for $p > 1$. However, for uniform convergence, we need some simpler conditions or tests like ordinary series.

3 Weirstrass M test-Test for convergence

Let $f_n(x)$ be a sequence of functions on A such that for every $n \geq 1$, $|f_n(x)| \leq M_n$ for any $x \in A$. Then the series $\sum_n f_n(x)$ converges absolutely and uniformly if the series $\sum_n M_n$ is convergent.

The above test only provides a sufficient condition and hence it may be possible that $\sum_n f_n(x)$ converges uniformly but $\sum_n M_n$ diverges for best possible choice of M_n. However, the best possible choice for M_n is given by

$$M_n = \sup_{x \in A} |f_n(x)|.$$

3.1 Examples

Example 1: Consider the series: $\sum_{n=1}^{\infty} Cos(nx)/n^p$, for $x \in R$. It is well known that $|Cos(nx)/n^p| \leq 1/n^p$ for all $n \geq 1$ and $x \in R$. Thus we define $M_n = 1/n^p$. Since the series $\sum_n M_n$ converges for $p > 1$, the given series converges absolutely and uniformly on $x \in R$.

Example 2: Consider the series: $\sum_{n=1}^{\infty} x^2/(1+n^2x^2)$, for $x \in R$. Now

$$x^2/(1+n^2x^2) = n^2x^2/(1+n^2x^2)n^{-2} < n^{-2} \ \forall x \in R.$$

Thus we take $M_n = 1/n^2$. Since the series $\sum_n M_n$ converges, the given series converges absolutely and uniformly on $x \in R$.

Example 3: Consider the series: $\sum_{n=1}^{\infty} x/(n+n^2x^2)$, for $x \in R$. Here $f_n(x) = x/(n+n^2x^2)$ and hence $|f_n(x)| = |x|/(n+n^2x^2)$ is a symmetric function about 0. Thus we consider only $x > 0$. Now a simple algebra shows that $x/(n+n^2x^2), x > 0$ attains maximum at $x = 1/\sqrt{n}$. Thus we take $M_n = 1/(2n^{3/2})$. Since $\sum_n M_n$ converges, the given series converges absolutely and uniformly on $x \in R$.

Example 4: Consider the series: $\sum_{n=1}^{\infty} x/(n+n^3x^2)$, for $x \in R$. Here $f_n(x) = x/(n+n^3x^2)$ and hence $|f_n(x)| = |x|/(n+n^3x^2)$ is a symmetric function about 0. Now it is well

known that $n + n^3x^2 \geq 2n^2|x|$, which implies that $|f_n(x)| \leq \frac{1}{2n^2}$ for $x \in R$. Thus we take $M_n = 1/(2n^2)$. Hence the given series converges absolutely and uniformly on $x \in R$.

4 Consequences of uniform convergence for series

We have already discussed different consequences of uniform convergence for sequence of functions regarding exchangeability of different mathematical operations like limits, differentiation and integration. Now we shall extend those ideas for a uniformly convergent series of functions.

4.1 Uniform convergence & continuity

Consider a series of functions $\sum_n f_n(x)$. Suppose

(i) $f_n(x)$ is continuous on A for every $n \geq 1$ and

(ii) $\sum f_n(x) \rightrightarrows f(x)$ on A.

Then $f(x)$ is continuous on A.

4.2 Interchangeability of summation & limit

Suppose $\sum f_n(x) \rightrightarrows f(x)$ on A. Let x_0 be any point in A. Then $\lim_{x \to x_0} f(x)$ exists and

$$\lim_{x \to x_0} \sum_n f_n(x) = \sum_n \lim_{x \to x_0} f_n(x).$$

Thus the operations of summation and limit are interchangeable under uniform convergence.

4.3 Interchangeability of differentiation & limit

Consider a series of functions $\sum_n f_n(x),\ x \in [a, b]$ such that

(i) $\sum_n f_n(x_0)$ converges to $f(x_0)$ for some $x_0 \in A$,

(ii) $f'_n(x)$ exists at all $x \in (a, b)$ and (iii) $\sum_n f'_n(x)$ converges uniformly on [a,b].

Then $\frac{d}{dx} \sum f_n(x) = \sum_n \frac{d}{dx} f_n(x)$.

4.4 Interchangeability of integration & limit

Consider a series of functions $\sum_n f_n(x),\ x \in [a,b]$ such that

(i) each $f_n(x)$ is Riemann integrable on [a,b] and

(ii) $\sum_n f_n(x)$ converges uniformly on [a,b].

Then $\sum_n \int_a^b f_n(x)dx = \int_a^b \sum_n f_n(x)dx$.

Large sample Inference: Module 9[1]

What we provide in this module

- Power series
- Convergence & applications
- Taylor series & applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Power series

Power series is a particular type of series of functions. Most of the discrete distributions like Binomial, Poisson, Negative binomial are actually a convergent power series. In addition, we use different generating functions, which are also power series. Thus existence of some generating function can be justified from the convergence properties of power series. We often deduce moments from the moment generating function using successive differentiation. Convergence of power series validates such operations.

1.1 Power series: Definition& Convergence

The series of functions $\sum_{n=0}^{\infty} f_n(x), x \in A$ is said to be a power series around c if $f_n(x) = a_n(x-c)^n, n \geq 0$. For simplicity we take c=0. Naturally a power series does not always converge for every x. However, it always converges to a_0 for $x = c$. For example, $\sum_{n=0}^{\infty} x^n/(n!)$ converges for every real x, whereas the series $\sum_{n=0}^{\infty} n!x^n$ converges only for x=0.

Now we shall discuss convergence of a power series. Consider the power series $\sum_{n=0}^{\infty} a_n(x-c)^n, x \in A$. Assume that one of the following limits exist(may be infinite)

(i) $\lim |a_{n+1}/a_n| = 1/R$

(ii) $\lim |a_n|^{1/n} = 1/R$

Then the power series converges absolutely for all x such that $|x - c| < R$. The power series diverges for all x such that $|x - c| > R$. In addition, the power series converges uniformly and absolutely for every compact subset of $(-R, R)$.

R as defined above is called the radius of convergence of the concerned power series. The interval $(-R, R)$ is called the interval of convergence of the power series. If R=0, we conclude that the power series is nowhere convergent. However, if $R = \infty$, the power series is said to be convergent everywhere. Every power series can be differentiated term by term within its interval of convergence. Every power series can also be integrated term by term within its interval of convergence.

1.2 Examples

Example 1: Consider the power series: $x + x^2 + x^3 + x^4 + ...$ Naturally this is the so called geometric series. However, one can alternatively interpret this as a power series with $a_n = 1 \; \forall \; n \geq 1$. Now $\lim |a_n|^{1/n} = 1$ and hence the power series converges absolutely for all $|x| < 1$. Thus the radius of convergence is $R = 1$ and interval of convergence (-1,1). It is already noted that the series diverges for $x = \pm 1$ and hence (-1,1) is the exact interval of convergence.

Example 2: Consider the power series: $\sum_{n=1}^{\infty} (-x)^n / n$. This is the so called log series as well as a power series with $a_n = (-1)^n / n \; \forall \; n \geq 1$. Now $\lim |a_n|^{1/n} = 1$ as $\lim n^{1/n} = 1$ and hence the power series converges absolutely for all $|x| < 1$. Thus the radius of convergence is $R = 1$ and interval of convergence (-1,1). However, for $x = -1$, we get the so called harmonic series which diverges. Again for $x = +1$ we get the alternating series which converges. Thus we get (-1,1] is the exact interval of convergence.

Example 3: Consider the power series: $\sum_{n=1}^{\infty} (x)^n / n!$. This is the well known e series and is also a power series with $a_n = 1/n! \; \forall \; n \geq 1$. Now $\lim |a_{n+1}/a_n| = 0$ and hence the power series converges absolutely for all $x \in (-\infty, \infty)$. Thus the radius of convergence in this case is $R = \infty$ and interval of convergence is the whole real line.

Example 4: Consider the power series: $\sum_{n=1}^{\infty} n!(x)^n$. This is a power series with $a_n = n! \; \forall \; n \geq 1$. Now $\lim |a_{n+1}/a_n| = \infty$ and hence the power series does not converge for any $x \in (-\infty, \infty)$ except for x=0. The radius of convergence in this case is $R = 0$.

Example 5: Consider the power series: $\sum_{n=2}^{\infty} (x+2)^n / log(n)$. This is a power series centered at c=-2 with $a_n = 1/log(n) \; \forall \; n \geq 2$. Now $\lim |a_{n+1}/a_n| = 1$ and hence the power series converges for any $|x + 2| < 1$ or for $-3 < x < -1$. The radius of convergence in this case is $R = 1$ with interval of convergence as (-3,-1). However, for x=-3, we get the series $\sum (-1)^n / log(n)$. Now, $1/log(n), n \geq 2$ is a decreasing sequence with limit zero. Hence by Leibnitz test the series converges. For x=-1, we get the series $\sum 1/log(n)$, which by

condensation test diverges. Hence the exact interval of convergence is $[-3, -1)$.

Example 6: Consider the power series: $\sum_{n=2}^{\infty} n!(x-1)^n/n^n$. This is a power series centered at c=1 with $a_n = n!/n^n \ \forall \ n \geq 2$. Now $\lim |a_{n+1}/a_n| = e$ and hence the power series converge for any x satisfying $|x-1| < e$ or for $-e+1 < x < e+1$. The radius of convergence in this case is $R = e$ with interval of convergence as (-e+1,e+1). However, separate examination on the boundary can be carried out to know the exact interval of convergence.

Example 7: Consider the hypergeometric series:

$$1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2.\gamma(\gamma+1)}x^2 + ...$$

with α, β, γ and x are all positive. This is a power series with $a_n = \frac{\alpha(\alpha+1)...(\alpha+n-1)\beta(\beta+1)...(\beta+n-1)}{1.2...n.\gamma.(\gamma+1)...(\gamma+n-1)}, n \geq 2$. Then we get $\lim |a_{n+1}/a_n| = 1$ and hence the power series converges for any x satisfying $|x| < 1$. The radius of convergence in this case is $R = 1$ with interval of convergence as (-1,1). However, convergence on the boundary $x = \pm 1$ depends on α, β and γ.

2 Taylor series expansion

Taylor series expansion is one of the most useful tools for a statistician. A Taylor series provides a power series representation of a function using its higher order derivatives. It is extremely useful for obtaining the large sample distribution of different implicit / explicit functions of relevant statistics. We start with the formal definition of a Taylor Series.

2.1 Taylor's Theorem: Univariate case

Let $f(x), x \in [a, b]$ be a function defined on the real line such that

(i) $f^{(k)}(x)$ is continuous on [a,b] for all $k \leq n-1$ and

(ii) $f^{(n)}(x)$ exists on [a,b].

Then for any $x_0 \in [a, b]$ we have

$$f(x) = P_n(x) + R_n(x),$$

where $P_n(x) = \sum_{k=0}^{n} f^{(k)}(x)(x - x_0)^k/k!$ with $f^{(0)}(x) = f(x)$ and $R_n(x) = f^{(n+1)}(c)(x - x_0)^{n+1}/(n+1)!$ where $c = tx + (1-t)x_0$ for $t \in [0, 1]$.

Note that P_n is called the n th Taylor polynomial for f at x_0 and it is a reasonable approximation to f for points near x_0. R_n is called the remainder term in Lagrange's form. If $\lim R_n(x) = 0$ for all $x \in (a, b)$, then $\lim P_n(x) = f(x) \forall\ x$. Thus f admits of the Taylor series expansion $f(x) = \sum_{k=0}^{\infty} f^{(k)}(x)(x - x_0)^k/k!$. If $x_0 = 0$, the above is called Maclaurin expansion of f(x).

However convergence is difficult to check for a Taylor series and hence we need some sufficient conditions. Consider the n th term, which is $f^{(n)}(x)(x - x_0)^n/n!$. Now $\sum_{n=0}^{\infty}(x - x_0)^n/n!$ converges for all x. Thus if we can show that $f^{(n)}(x)$ is bounded by some finite number M for all x in some interval around x_0, the convergence can be asserted for that interval.

2.1.1 Examples

Example 1: Consider $f(x) = sin(x)$ and assume $x_0 = 0$. Then $f^{(n)}(x) = Sin(n\pi/2 + x)$ Thus $R_n(x) = Sin(\overline{n+1}\pi/2 + c)x^{n+1}/(n+1)!$. Now it is easy to observe that $|Sin(n\pi/2 + x)| < 1$ for any $n \geq 1$, and hence by the condition defined above, the Taylor series converges to the sine function on the real line. The same is valid for cosine function.

Example 2: Consider $f(x) = log(1 + x)$ and assume $x_0 = 0$. Since $f^{(n)}(x) = (-1)^{n-1}(n-1)!(1+x)^{-n}$, the Taylor series around 0 is $\sum_{n=0}^{\infty}(-1)^{n-1}x^n/n$. Suppose we wish to know whether the above series converges for x=1. Now $R_n(x) = (-1)^n(n)!(1 + c)^{-n-1}x^{n+1}/(n+1)!, c \in (0, 1)$. Since $|R_n(1)| \leq 1/n \to 0$, the series converges. Thus we get the representation $log(2) = \sum_{n=1}^{\infty}(-1)^{n-1}/n$.

2.2 Taylor's Theorem: Multivariate case

Let $f = f(x_1, x_2, .., x_s)$ be a function of s variables defined on $A \in R^s$. Define the gradient operator as $\nabla = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_s})'$. Also define the partial differential operator $x'\nabla = \sum_{i=1}^{s} x_i \frac{\partial}{\partial x_i}$. The operator $(x'\nabla)^k$ is used as the k th power of a sum.

Multivariate Taylor's Theorem: First order: Let $f = f(x_1, x_2, .., x_s)$ be a function of s variables defined on $A \subseteq R^s$. Suppose there exists a neighbourhood of a in A such that partial derivatives up to order $(l+1)$ are continuous in the neighbourhood. Then for any x belonging to that neighbourhood,

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \frac{\{(x-a)'\nabla\}^{k+1}}{(k+1)!} f(z),$$

where $z = ta + (1-t)x$ for some $t \in [0, 1]$.

Multivariate Taylor's Theorem: Second order: However, the second order expansion is often expressed in terms of Hessian matrix $H(x) = \frac{\partial^2 f(x)}{\partial x x'}$ as $f(x) = f(a) + \frac{\partial f(x)}{\partial x'}|_{x=a} + \frac{1}{2}(x-a)'H(z)(x-a)$, where $z = ta + (1-t)x$ for some $t \in [0, 1]$.

Multivariate Taylor's Theorem using $\bigcirc$ and $\circ$: If all the partial derivatives of f up to order $(l+1)$ are bounded in the neighbourhood of a, we have either

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \circ(|x-a|^l),$$

or

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \bigcirc(|x-a|^{l+1}),$$

where $|x-a| = \sqrt{\sum_{i=1}^{s}(x_i - a)^2}$.

Large sample Inference: Module 8[1]

What we provide in this module

- Consequences of uniform convergence for sequence of functions
- Series of functions & their convergence
- Examples
- Consequences of uniform convergence for series of functions

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Consequences of uniform convergence

Now we shall discuss few consequences of uniform convergence. In fact, uniform convergence allows exchangeability of different operations like limits, differentiation, integration, etc.

1.1 Uniform convergence & limit interchanging

If $f_n(x) \rightrightarrows f(x)$ on A, where f_n is continuous on A for every n. Then f is also continuous on A. Thus for any $a \in A$, we have

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty}\lim_{x\to a} f_n(x).$$

The implication of the above is that if the limits are not interchangeable, then uniform convergence does not hold. For example, consider $f_n(x) = x^n, x \in [0, 1)$ then the pointwise limit is the function $f(x) = 0, x \in [0, 1)$. However $\lim_{n\to\infty} f_n(x) = 0$. Thus $\lim_{x\to 1} f(x) \neq \lim_{n\to\infty}\lim_{x\to a} f_n(x)$. Hence the convergence can not be uniform.

1.2 Uniform convergence & differentiation

If $f_n(x)$ be a sequence of functions on [a,b] such that

(i) for some $x_0 \in (a, b)$, $f_n(x_0) \to f(x_0)$ and

(ii) $f_n'(x)$ exists at all $x \in [a, b]$ and converges uniformly to some function g on [a,b].

Then $f_n(x) \rightrightarrows f(x)$ on [a,b] and $g(x) = f'(x)\ \forall\ x \in [a, b]$. That is

$$\lim_{n\to\infty} \frac{df_n(x)}{dx} = \frac{d}{dx}\lim_{n\to\infty} f_n(x)\ \forall\ x \in [a, b].$$

For better understanding, consider an example with $f_n(x) = \frac{x}{1+n^2x^2}, |x| \leq 1$. Define $f(x) = 0, x \in [-1, 1]$, then it can be shown that $M_n = \sup_{x\in[-1,1]} |f_n(x) - f(x)| = 1/2n \to 0$ as $n \to \infty$. Thus the convergence is uniform. However $f_n'(x) = \frac{1-n^2x^2}{(1+n^2x^2)^2}$ and hence for $x = 0$, $\lim_{n\to\infty} f_n'(x) = 1$. Thus $\lim_{n\to\infty} f_n'(x) \neq f'(x)$ for x=0.

1.3 Uniform convergence & integration

Suppose $f_n(x)$ is a sequence of Riemann integrable functions on [a,b] such that $f_n(x) \rightrightarrows f(x)$ on [a,b]. Then f is also Riemann integrable on [a,b] and

$$\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b \lim_{n\to\infty} f_n(x)dx.$$

For an example, consider $f_n(x) = nxexp(-nx^2), x \in [0,1]$. Define $f(x) = 0, x \in [0,1]$, then it can be shown that $M_n = \sup_{x\in[0,1]} |f_n(x) - f(x)| = \sqrt{n/2} \to \infty$ as $n \to \infty$. Thus the convergence is not uniform. Thus $\int_0^1 f_n(x)dx = (1 - exp(-n))/2 \to 1/2$ but $\int_0^1 f(x)dx = 0$. However if we consider the interval [1,3], then the convergence is uniform and hence $\int_1^3 f_n(x)dx \to 0 = \int_1^3 f(x)dx$.

2 Series of functions-Motivation

We have already discussed sequence of functions and their convergence. Like ordinary series, the idea of sequence of functions can also be extended to series of functions. For example, we have well known e series and log series. As earlier, we can have different concepts of convergence.

Let $\{f_n(x)\}$ be a sequence of functions on A to R. Then $f_1+f_2+f_3+..$ is said to be a series or infinite series and is often denoted by $\sum_{n=1}^{\infty} f_n(x)$. Define the sequence of partial sums of the infinite series by $s_n(x) = \sum_{k=1}^{n} f_k(x)$. If the sequence $\{s_n(x)\}$ converges pointwise to some $s(x)$ on A, the series $\sum_{n=1}^{\infty} f_n(x)$ is said to converge pointwise to $s(x)$ on A. However, if the convergence is uniform to some $s(x)$ on A, the series $\sum_{n=1}^{\infty} f_n(x)$ converges uniformly to $s(x)$ on A. f the series $\sum_{n=1}^{\infty} |f_n(x)|$ converges for each $x \in A$, $\sum_{n=1}^{\infty} f_n(x)$ is said to converge absolutely on A.

2.1 Examples

Example 1: Consider the series: $x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + ...$, for $x \in [0,1]$. Then $s_n(x) = \sum_{i=0}^{n-1} \frac{x^2}{(1+x^2)^i} = (1+x^2) - (1+x^2)^{-n+1}$ is the n th partial sum. Thus we find that

$$\begin{aligned} \lim s_n(x) &= 0 \text{ if } x = 0 \\ &= 1 + x^2 \text{ if } x \in (0,1] \end{aligned}$$

Thus $s_n(x)$ converges pointwise to $s(x) = (1+x^2)I(0 < x \leq 1)$ with I as an indicator function. Hence the series converges pointwise to $s(x)$ on [0,1]. However, s(x) is not continuous on [0,1] but each $s_n(x)$ is continuous on [0,1] and hence the convergence is not uniform on [0,1].

Example 2: Consider the series: $\sum_{n=1}^{\infty} x^n(1-x)^n$, for $x \in [0,1]$. Then $s_n(x) = \sum_{i=1}^{n} x^i(1-x)^i = x(1-x)\frac{1-[x(1-x)]^{n-1}}{1-x(1-x)}$ is the n th partial sum. Thus we find that

$$\begin{aligned} s(x) = \lim s_n(x) &= 0 \text{ if } x = 0,1 \\ &= \frac{x(1-x)}{1-x(1-x)} \text{ if } x \in (0,1) \end{aligned}$$

Then $s_n(x)$ converges pointwise to $s(x)$ and hence the series converges pointwise to $s(x)$ on [0,1].

Example 3: Consider the series: $\sum_{n=1}^{\infty} Sin(nx)/n^p$, for $x \in R$. Let us fix x at a and consider the convergence of $\sum_n Sin(na)/n^p$. Now $|Sin(na)/n^p| \leq 1/n^p$ for all $n \geq 1$. Hence by comparison test $\sum_n |Sin(na)|/n^p$ converges for $p > 1$, that is the series converges absolutely. Since a is arbitrary, the series $\sum_{n=1}^{\infty} Sin(nx)/n^p$ is absolutely convergent on R for $p > 1$. However, for uniform convergence, we need some simpler conditions or tests like ordinary series.

3 Weirstrass M test-Test for convergence

Let $f_n(x)$ be a sequence of functions on A such that for every $n \geq 1$, $|f_n(x)| \leq M_n$ for any $x \in A$. Then the series $\sum_n f_n(x)$ converges absolutely and uniformly if the series $\sum_n M_n$ is convergent.

The above test only provides a sufficient condition and hence it may be possible that $\sum_n f_n(x)$ converges uniformly but $\sum_n M_n$ diverges for best possible choice of M_n. However, the best possible choice for M_n is given by

$$M_n = \sup_{x \in A} |f_n(x)|.$$

3.1 Examples

Example 1: Consider the series: $\sum_{n=1}^{\infty} Cos(nx)/n^p$, for $x \in R$. It is well known that $|Cos(nx)/n^p| \leq 1/n^p$ for all $n \geq 1$ and $x \in R$. Thus we define $M_n = 1/n^p$. Since the series $\sum_n M_n$ converges for $p > 1$, the given series converges absolutely and uniformly on $x \in R$.

Example 2: Consider the series: $\sum_{n=1}^{\infty} x^2/(1+n^2x^2)$, for $x \in R$. Now

$$x^2/(1+n^2x^2) = n^2x^2/(1+n^2x^2)n^{-2} < n^{-2} \ \forall x \in R.$$

Thus we take $M_n = 1/n^2$. Since the series $\sum_n M_n$ converges, the given series converges absolutely and uniformly on $x \in R$.

Example 3: Consider the series: $\sum_{n=1}^{\infty} x/(n+n^2x^2)$, for $x \in R$. Here $f_n(x) = x/(n+n^2x^2)$ and hence $|f_n(x)| = |x|/(n+n^2x^2)$ is a symmetric function about 0. Thus we consider only $x > 0$. Now a simple algebra shows that $x/(n+n^2x^2), x > 0$ attains maximum at $x = 1/\sqrt{n}$. Thus we take $M_n = 1/(2n^{3/2})$. Since $\sum_n M_n$ converges, the given series converges absolutely and uniformly on $x \in R$.

Example 4: Consider the series: $\sum_{n=1}^{\infty} x/(n+n^3x^2)$, for $x \in R$. Here $f_n(x) = x/(n+n^3x^2)$ and hence $|f_n(x)| = |x|/(n+n^3x^2)$ is a symmetric function about 0. Now it is well

known that $n + n^3x^2 \geq 2n^2|x|$, which implies that $|f_n(x)| \leq \frac{1}{2n^2}$ for $x \in R$. Thus we take $M_n = 1/(2n^2)$. Hence the given series converges absolutely and uniformly on $x \in R$.

4 Consequences of uniform convergence for series

We have already discussed different consequences of uniform convergence for sequence of functions regarding exchangeability of different mathematical operations like limits, differentiation and integration. Now we shall extend those ideas for a uniformly convergent series of functions.

4.1 Uniform convergence & continuity

Consider a series of functions $\sum_n f_n(x)$. Suppose
(i) $f_n(x)$ is continuous on A for every $n \geq 1$ and
(ii) $\sum f_n(x) \rightrightarrows f(x)$ on A.
Then $f(x)$ is continuous on A.

4.2 Interchangeability of summation & limit

Suppose $\sum f_n(x) \rightrightarrows f(x)$ on A. Let x_0 be any point in A. Then $\lim_{x\to x_0} f(x)$ exists and

$$\lim_{x\to x_0} \sum_n f_n(x) = \sum_n \lim_{x\to x_0} f_n(x).$$

Thus the operations of summation and limit are interchangeable under uniform convergence.

4.3 Interchangeability of differentiation & limit

Consider a series of functions $\sum_n f_n(x),\ x \in [a, b]$ such that
(i) $\sum_n f_n(x_0)$ converges to $f(x_0)$ for some $x_0 \in A$,
(ii) $f_n'(x)$ exists at all $x \in (a, b)$ and (iii) $\sum_n f_n'(x)$ converges uniformly on [a,b].
Then $\frac{d}{dx} \sum f_n(x) = \sum_n \frac{d}{dx} f_n(x)$.

4.4 Interchangeability of integration & limit

Consider a series of functions $\sum_n f_n(x),\ x \in [a,b]$ such that

(i) each $f_n(x)$ is Riemann integrable on [a,b] and

(ii) $\sum_n f_n(x)$ converges uniformly on [a,b].

Then $\sum_n \int_a^b f_n(x)dx = \int_a^b \sum_n f_n(x)dx$.

Large sample Inference: Module 9[1]

What we provide in this module

- Power series
- Convergence & applications
- Taylor series & applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Power series

Power series is a particular type of series of functions. Most of the discrete distributions like Binomial, Poisson, Negative binomial are actually a convergent power series. In addition, we use different generating functions, which are also power series. Thus existence of some generating function can be justified from the convergence properties of power series. We often deduce moments from the moment generating function using successive differentiation. Convergence of power series validates such operations.

1.1 Power series: Definition& Convergence

The series of functions $\sum_{n=0}^{\infty} f_n(x), x \in A$ is said to be a power series around c if $f_n(x) = a_n(x-c)^n, n \geq 0$. For simplicity we take c=0. Naturally a power series does not always converge for every x. However, it always converges to a_0 for $x = c$. For example, $\sum_{n=0}^{\infty} x^n/(n!)$ converges for every real x, whereas the series $\sum_{n=0}^{\infty} n!x^n$ converges only for x=0.

Now we shall discuss convergence of a power series. Consider the power series $\sum_{n=0}^{\infty} a_n(x-c)^n, x \in A$. Assume that one of the following limits exist(may be infinite)
(i) $\lim |a_{n+1}/a_n| = 1/R$
(ii) $\lim |a_n|^{1/n} = 1/R$
Then the power series converges absolutely for all x such that $|x-c| < R$. The power series diverges for all x such that $|x-c| > R$. In addition, the power series converges uniformly and absolutely for every compact subset of $(-R, R)$.
R as defined above is called the radius of convergence of the concerned power series. The interval $(-R, R)$ is called the interval of convergence of the power series. If R=0, we conclude that the power series is nowhere convergent. However, if $R = \infty$, the power series is said to be convergent everywhere. Every power series can be differentiated term by term within its interval of convergence. Every power series can also be integrated term by term within its interval of convergence.

1.2 Examples

Example 1: Consider the power series: $x + x^2 + x^3 + x^4 + ...$ Naturally this is the so called geometric series. However, one can alternatively interpret this as a power series with $a_n = 1 \ \forall \ n \geq 1$. Now $\lim |a_n|^{1/n} = 1$ and hence the power series converges absolutely for all $|x| < 1$. Thus the radius of convergence is $R = 1$ and interval of convergence (-1,1). It is already noted that the series diverges for $x = \pm 1$ and hence (-1,1) is the exact interval of convergence.

Example 2: Consider the power series: $\sum_{n=1}^{\infty}(-x)^n/n$. This is the so called log series as well as a power series with $a_n = (-1)^n/n \ \forall \ n \geq 1$. Now $\lim |a_n|^{1/n} = 1$ as $\lim n^{1/n} = 1$ and hence the power series converges absolutely for all $|x| < 1$. Thus the radius of convergence is $R = 1$ and interval of convergence (-1,1). However, for $x = -1$, we get the so called harmonic series which diverges. Again for $x = +1$ we get the alternating series which converges. Thus we get (-1,1] is the exact interval of convergence.

Example 3: Consider the power series: $\sum_{n=1}^{\infty}(x)^n/n!$. This is the well known e series and is also a power series with $a_n = 1/n! \ \forall \ n \geq 1$. Now $\lim |a_{n+1}/a_n| = 0$ and hence the power series converges absolutely for all $x \in (-\infty, \infty)$. Thus the radius of convergence in this case is $R = \infty$ and interval of convergence is the whole real line.

Example 4: Consider the power series: $\sum_{n=1}^{\infty} n!(x)^n$. This is a power series with $a_n = n! \ \forall \ n \geq 1$. Now $\lim |a_{n+1}/a_n| = \infty$ and hence the power series does not converge for any $x \in (-\infty, \infty)$ except for x=0. The radius of convergence in this case is $R = 0$.

Example 5: Consider the power series: $\sum_{n=2}^{\infty}(x + 2)^n/log(n)$. This is a power series centered at c=-2 with $a_n = 1/log(n) \ \forall \ n \geq 2$. Now $\lim |a_{n+1}/a_n| = 1$ and hence the power series converges for any $|x + 2| < 1$ or for $-3 < x < -1$. The radius of convergence in this case is $R = 1$ with interval of convergence as (-3,-1). However, for x=-3, we get the series $\sum(-1)^n/log(n)$. Now, $1/log(n), n \geq 2$ is a decreasing sequence with limit zero. Hence by Leibnitz test the series converges. For x=-1, we get the series $\sum 1/log(n)$, which by

condensation test diverges. Hence the exact interval of convergence is $[-3,-1)$.

Example 6: Consider the power series: $\sum_{n=2}^{\infty} n!(x-1)^n/n^n$. This is a power series centered at c=1 with $a_n = n!/n^n \ \forall \ n \geq 2$. Now $\lim |a_{n+1}/a_n| = e$ and hence the power series converge for any x satisfying $|x-1| < e$ or for $-e+1 < x < e+1$. The radius of convergence in this case is $R = e$ with interval of convergence as (-e+1,e+1). However, separate examination on the boundary can be carried out to know the exact interval of convergence.

Example 7: Consider the hypergeometric series:

$$1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2.\gamma(\gamma+1)}x^2 + ...$$

with α, β, γ and x are all positive. This is a power series with $a_n = \frac{\alpha(\alpha+1)...(\alpha+n-1)\beta(\beta+1)...(\beta+n-1)}{1.2...n.\gamma.(\gamma+1)...(\gamma+n-1)}, n \geq 2$. Then we get $\lim |a_{n+1}/a_n| = 1$ and hence the power series converges for any x satisfying $|x| < 1$. The radius of convergence in this case is $R = 1$ with interval of convergence as (-1,1). However, convergence on the boundary $x = \pm 1$ depends on α, β and γ.

2 Taylor series expansion

Taylor series expansion is one of the most useful tools for a statistician. A Taylor series provides a power series representation of a function using its higher order derivatives. It is extremely useful for obtaining the large sample distribution of different implicit / explicit functions of relevant statistics. We start with the formal definition of a Taylor Series.

2.1 Taylor's Theorem: Univariate case

Let $f(x), x \in [a,b]$ be a function defined on the real line such that
(i) $f^{(k)}(x)$ is continuous on [a,b] for all $k \leq n-1$ and
(ii) $f^{(n)}(x)$ exists on [a,b].

Then for any $x_0 \in [a,b]$ we have

$$f(x) = P_n(x) + R_n(x),$$

where $P_n(x) = \sum_{k=0}^{n} f^{(k)}(x)(x-x_0)^k/k!$ with $f^{(0)}(x) = f(x)$ and $R_n(x) = f^{(n+1)}(c)(x-x_0)^{n+1}/(n+1)!$ where $c = tx + (1-t)x_0$ for $t \in [0,1]$.

Note that P_n is called the n th Taylor polynomial for f at x_0 and it is a reasonable approximation to f for points near x_0. R_n is called the remainder term in Lagrange's form. If $\lim R_n(x) = 0$ for all $x \in (a,b)$, then $\lim P_n(x) = f(x) \forall\ x$. Thus f admits of the Taylor series expansion $f(x) = \sum_{k=0}^{\infty} f^{(k)}(x)(x-x_0)^k/k!$. If $x_0 = 0$, the above is called Maclaurin expansion of f(x).

However convergence is difficult to check for a Taylor series and hence we need some sufficient conditions. Consider the n th term, which is $f^{(n)}(x)(x-x_0)^n/n!$. Now $\sum_{n=0}^{\infty}(x-x_0)^n/n!$ converges for all x. Thus if we can show that $f^{(n)}(x)$ is bounded by some finite number M for all x in some interval around x_0, the convergence can be asserted for that interval.

2.1.1 Examples

Example 1: Consider $f(x) = sin(x)$ and assume $x_0 = 0$. Then $f^{(n)}(x) = Sin(n\pi/2 + x)$ Thus $R_n(x) = Sin(\overline{n+1}\pi/2 + c)x^{n+1}/(n+1)!$. Now it is easy to observe that $|Sin(n\pi/2 + x)| < 1$ for any $n \geq 1$, and hence by the condition defined above, the Taylor series converges to the sine function on the real line. The same is valid for cosine function.

Example 2: Consider $f(x) = log(1+x)$ and assume $x_0 = 0$. Since $f^{(n)}(x) = (-1)^{n-1}(n-1)!(1+x)^{-n}$, the Taylor series around 0 is $\sum_{n=0}^{\infty}(-1)^{n-1}x^n/n$. Suppose we wish to know whether the above series converges for x=1. Now $R_n(x) = (-1)^n(n)!(1+c)^{-n-1}x^{n+1}/(n+1)!, c \in (0,1)$. Since $|R_n(1)| \leq 1/n \to 0$, the series converges. Thus we get the representation $log(2) = \sum_{n=1}^{\infty}(-1)^{n-1}/n$.

2.2 Taylor's Theorem: Multivariate case

Let $f = f(x_1, x_2, .., x_s)$ be a function of s variables defined on $A \in R^s$. Define the gradient operator as $\nabla = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_s})'$. Also define the partial differential operator $x'\nabla = \sum_{i=1}^{s} x_i \frac{\partial}{\partial x_i}$. The operator $(x'\nabla)^k$ is used as the k th power of a sum.

Multivariate Taylor's Theorem: First order: Let $f = f(x_1, x_2, .., x_s)$ be a function of s variables defined on $A \subseteq R^s$. Suppose there exists a neighbourhood of a in A such that partial derivatives up to order $(l+1)$ are continuous in the neighbourhood. Then for any x belonging to that neighbourhood,

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \frac{\{(x-a)'\nabla\}^{k+1}}{(k+1)!} f(z),$$

where $z = ta + (1-t)x$ for some $t \in [0, 1]$.

Multivariate Taylor's Theorem: Second order: However, the second order expansion is often expressed in terms of Hessian matrix $H(x) = \frac{\partial^2 f(x)}{\partial x x'}$ as $f(x) = f(a) + \frac{\partial f(x)}{\partial x'}|_{x=a} + \frac{1}{2}(x-a)'H(z)(x-a)$, where $z = ta + (1-t)x$ for some $t \in [0, 1]$.

Multivariate Taylor's Theorem using $\bigcirc$ and $\circ$: If all the partial derivatives of f up to order $(l+1)$ are bounded in the neighbourhood of a, we have either

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \circ(|x-a|^l),$$

or

$$f(x) - f(a) = \sum_{k=1}^{l} \frac{\{(x-a)'\nabla\}^k}{k!} f(a) + \bigcirc(|x-a|^{l+1}),$$

where $|x-a| = \sqrt{\sum_{i=1}^{s}(x_i - a)^2}$.

Large sample Inference: Module 10[1]

What we provide in this module

- Convergence in probability
- Convergence in distributions
- Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Convergence Of A Sequence Of Random Variables

Convergence of a sequence of random variables is the central concept in Asymptotic Theory. Two types of approximations are of great importance in statistics. The first type approximates a given distribution function by the another. In the second type, a given random variable is approximated by another random variable.
Corresponding to the first type, we have convergence in distributions. However, connected with the latter type, we identify three distinct types of convergence,namely

convergence in probability

almost sure convergence

convergence in the r th mean.

2 Convergence in distributions

2.1 Basic ideas

The pivotal quantity in convergence in distributions, is the idea of a distribution function. So we provide some basics related to distribution functions. For a random variable X, its cdf is defined as $F(x) = P(X \leq x), x \in R$. Then it is well known that

F is non-decreasing

F is right continuous, i.e. $F(x + 0) = F(x) \forall x$

$F(\infty) = 1$ and $F(-\infty) = 0$

The last condition ensures that F is proper.
Let X_n be a sequence of random variables. Suppose F_n is the corresponding sequence of distribution functions, i.e. $X_n \sim F_n(x)$. We want to investigate the convergence of F_n to some limiting cdf. Consider an example, where $P(X_n = n) = 1 \forall n$. Then $F_n(x) = I(x \geq n)$,

which is a sequence of functions. Thus the pointwise limit is simply: $F_n(x) \to 0 \forall\ x$. Hence the pointwise limit is not a proper cdf.
However, for a statistician, only convergence to a cdf is of interest and hence we should define convergence such that $F_n(x) \to F(x)$, for a cdf F(x), which may not be proper.

2.2 Weak convergence & convergence in distributions

Consider a sequence of distribution functions $F_n(x)$. Then $F_n \stackrel{w}{\to} F$, (i.e. converges weakly to F) if $F_n(x) \to F(x)$ for all x, where F(x) is continuous.
For weak convergence, we do not need a proper cdf F. If in addition, F is proper, we have convergence in distribution.
Consider a sequence of random variables X_n and distribution functions $F_n(x)$ with $X_n \sim F_n$. Then $X_n \stackrel{D}{\to} X$, (i.e. converges in distribution or law) if there exists a random variable X with cdf F(x) such that $F_n(x) \to F(x)$ for all x, where F(x) is continuous.

2.3 Examples

Example 1: Suppose $P(X_n = a + 1/n) = 1 \forall n \geq 1$. Then one would expect convergence to a random variable X such that $P(X = a) = 1$. However, $F_n(x) = I(x \geq a + 1/n)$. Thus $\lim F_n(x) = I(x \geq a)$ Hence, if we define $F(x) = I(x \geq a)$, we find $\lim F_n(a) = 0$ but $F(a) = 1$. Naturally, if we do not restrict to continuity points of F, convergence is not possible. However, a is not a continuity point of F and hence according to the definition $X_n \stackrel{D}{\to} X$.

Example 2: Suppose $X_n \sim N(0, 1 + 1/n^2) \forall n \geq 1$. Then we could expect that X_n converges to a standard normal variable. For better understanding, we plot the distribution of X_n for different values of n.

Naturally, we find that for increasing values of n, the graph approaches to the density of

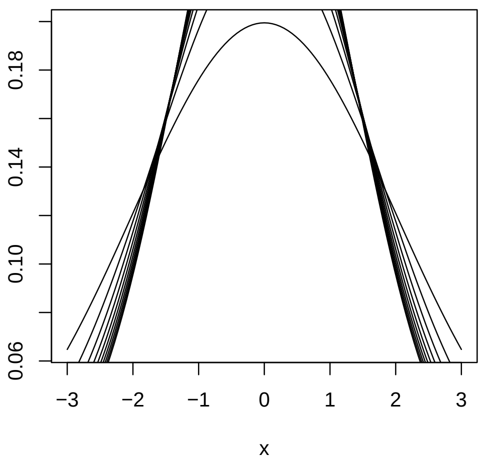


Figure 1: **Convergence in distribution**

a standard normal variable. However, $F_n(x) = \Phi(\frac{nx}{\sqrt{1+n^2}})$. Since, $\frac{nx}{\sqrt{1+n^2}} \to x$ as $n \to \infty$, we have $F_n(x) \to \Phi(x)$ pointwise on R. Hence, X_n converges to a standard normal variable.

Example 3: Suppose $X_n \sim N(0, n^2) \forall n \geq 1$. Then we have $F_n(x) = \Phi(x/n) \forall x$, so that $\lim F_n(x) = \frac{1}{2}$. Thus, if we define $F(x) = \frac{1}{2}$, we find $F(\infty) = 1/2 \neq 1$. Hence F is not proper so that $F_n \stackrel{w}{\to} F$. Naturally, we can't identify any rv X having df F and hence we do not have convergence in distributions.

Example 4: Suppose $X_n \sim N(0, 1/n^2) \forall n \geq 1$. Then we have $F_n(x) = \Phi(nx) \forall x$, so that $\lim F_n(x) = I(x > 0) + \frac{1}{2} I(x = 0)$. Thus, if we define $F(x) = I(x > 0) + \frac{1}{2} I(x = 0)$, we find that F is not right continuous at 0. But 0 is a point of discontinuity of F. Thus $X_n \stackrel{d}{\to} X$, where $P(X = 0) = 1$. The above can be justified from the graph in the next page.

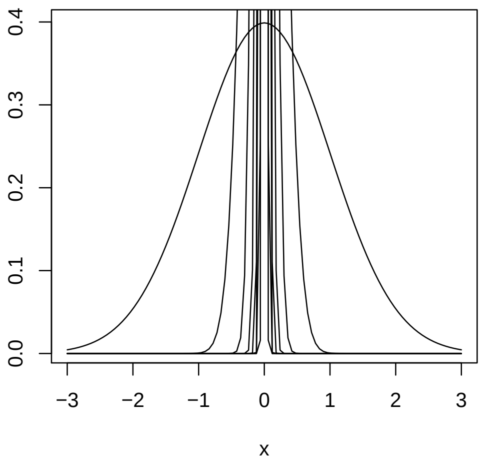


Figure 2: **Convergence in distribution**

2.4 Convergence in distributions: Uniform Convergence

The definitions of convergence in distribution is based on the concept of pointwise convergence of sequence of distribution functions. However, bounded and non-decreasing property of distribution function enables uniform convergence. The following theorem ensures uniform convergence.

Polya Theorem: Consider a sequence of distribution functions F_n and another cdf F(x). If $F_n(x) \to F(x)$ for all x, where F(x) is continuous then

$$\sup_x |F_n(x) - F(x)| \to 0$$

as $n \to \infty$.

3 Convergence in probability

For an ordinary sequence $a_n, n \geq 1$ converging to a, it is already noted that for given $\epsilon > 0$, $|a_n - a| < \epsilon$ is satisfied for large n. However, for a sequence of random variables, such an

inequality may not hold with probability one. But we can made the corresponding probability arbitrarily close to unity for appropriately chosen limit. In particular the corresponding probability sequence is an ordinary sequence and the above suggests that it converges to unity provided the limit random variable is chosen appropriately.

Convergence in probability: Consider a sequence of random variables X_n. Then X_n is said to converge in probability to another random variable X if for every given $\epsilon > 0$,

$$\lim P(d(X_n, X) \leq \epsilon) = 1 \Leftrightarrow \lim P(d(X_n, X) > \epsilon) = 0,$$

where $d(X_n, X) = |X_n - X|$ is the distance function and the random variables $X_n, n \geq 1$ and X are defined on a common sample space. For random vectors, one can use the Euclidean norm

$$d(X_n, X) = \parallel X_n - X \parallel = \sqrt{(X_n - X)'(X_n - X)}$$

.

We use the notation $X_n - X \xrightarrow{P} 0$, for stochastic X and $X_n \xrightarrow{P} X$ for nonrandom X.

3.1 An example

Consider $X_n = X + \frac{1}{n^2}$, where $X \sim N(0,1)$. Intuitively, it appears that for large n, the contribution of the second quantity is negligible and hence X_n is very close to X. Now $X_n - X$ is degenerate at $1/n^2$. Then for any $\epsilon > 1$, $P(|X_n - X| > \epsilon) = 0$. Again $P(|X_n - X| > \epsilon) = 1/n^2$ or 0 as $\epsilon < 1/n^2$ or $\epsilon \geq 1/n^2$. Thus for any $\epsilon > 0$, and $P(|X_n - X| > \epsilon) \to 0$ and hence $X_n \xrightarrow{P} X$

3.2 Convergence in probability & distribution

Convergence in probability implies convergence in distribution but the converse is true only for degenerate limits. As an example consider a random variable X having a N(0,1) distribution and define $X_n = (-1)^n X$. Then due to symmetry $X_n \stackrel{D}{=} N(0,1)$ and hence

$X_n \xrightarrow{D} Y = N(0,1)$. But the distribution of $|X_n - Y|$ is independent of any n and hence $P(|X_n - Y| > \epsilon)$ does not tend to zero. Thus convergence in probability does not hold though convergence in distribution holds.

Large sample Inference: Module 11[1]

What we provide in this module

- Almost sure convergence
- Weak law of large numbers
- Strong law of large numbers
- Borel-Cantelli Lemma

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Convergence almost surely

Convergence in probability does not ensure $\lim X_n = X$ with probability 1 and hence we have another mode of convergence, known as almost sure convergence.

Almost sure convergence: Consider a sequence of random variables X_n. Then X_n is said to converge almost surely to another random variable X . i.e. $X_n - X \overset{a.s.}{\to} 0$ if

$$\sup_{m \geq n} d(X_m, X) \overset{P}{\to} 0,$$

where the random variables $X_n, n \geq 1$ and X are defined on a common sample space. This is a stronger mode of convergence because $\sup_{m \geq n} d(X_m, X) \overset{P}{\to} 0 \Rightarrow d(X_n, X) \overset{P}{\to} 0$.

1.1 An example

Consider $X_n \overset{iid}{\sim} R(0,1), n \geq 1$. Then it is well known that $E(X_{(1:n)}) = \frac{1}{n+1} \to 0$ for large n. Now consider the quantity $P(|X_{(1:m)} - 0| \leq \epsilon \; \forall \; m \geq n)$. Since $X_{(1:m)} \in [0,1]$ with probability one, we get $|X_{(1:m)} - 0| \in [0,1]$ with probability one. Then for $\epsilon \geq 1$, the above probability is simply one. Now assume $0 < \epsilon < 1$, then the above probability reduces to $P(X_{(1:m)} \leq \epsilon \; \forall \; m \geq n)$. Since, $X_{(1:n)} \geq X_{(1:n+1)}$, we have $P(X_{(1:m)} \leq \epsilon \; \forall \; m \geq n) = P(X_{(1:m)} \leq \epsilon)$. Now $P(X_{(1:m)} \leq \epsilon) = 1 - P(X_{(1:m)} > \epsilon) = 1 - (1-\epsilon)^m \to 1$ and hence $X_n \overset{a.s.}{\to} 0$.

2 Convergence in the r th mean

Consider the random variables X_n and X defined on a common sample space such that $E|X_n|^r < \infty$ for some $r > 0$. Then X_n is said to converge in the r th mean to X if

$$E|X_n - X|^r \to 0$$

provided $E|X|^r < \infty$. In notation, we use $X_n \overset{r}{\to} X$. It should be noted that if $X_n \overset{r}{\to} X$ for some $r > 0$ then $X_n \overset{P}{\to} X$. However, the converse may not be always true.

3 Stochastic order relations

We have already introduced big O and small O notations for ordinary sequences. Similar notations can also be introduced to simplify the notions regarding stochastic convergence.

Stochastic big O: For a sequence of random variables X_n, if for every $\epsilon > 0$, there exists a positive constant $K(\epsilon)$ and an integer $n_0 = n_0(\epsilon)$ such that

$$P(|X_n| < K(\epsilon)) \geq 1 - \epsilon, \quad n \geq n_0(\epsilon),$$

then $X_n = O_p(1)$,i.e. X_n is bounded in probability.
For sequences of random variables X_n and Y_n, if for every $\epsilon > 0$, there exists a positive constant $K(\epsilon)$ and an integer $n_0 = n_0(\epsilon)$ such that

$$P(|\frac{X_n}{Y_n}| < K(\epsilon)) \geq 1 - \epsilon, \quad n \geq n_0(\epsilon),$$

then $X_n = O_p(Y_n)$.

Stochastic small o: For a sequence of random variables X_n, if for every $\epsilon > 0$ and $\eta > 0$, there exists a positive integer $n_0 = n_0(\epsilon, \eta)$ such that

$$P(|X_n| > \eta) < \epsilon, \quad n \geq n_0(\epsilon, \eta),$$

then $X_n = o_p(1)$.
It is easy to observe that $X_n = o_p(1) \Leftrightarrow X_n \xrightarrow{P} 0$. One can use Euclidean norm to extend the idea for vector valued random variables.
For sequences of random variables X_n and Y_n, if for every $\epsilon > 0$ and $\eta > 0$, there exists a positive integer $n_0 = n_0(\epsilon, \eta)$ such that

$$P(|\frac{X_n}{Y_n}| > \eta) < \epsilon, \quad n \geq n_0(\epsilon, \eta),$$

then $X_n = o_p(Y_n)$.

For a sequence of random variables X_n, if $X_n = o_p(1)$ then $X_n = O_p(1)$ also but the converse is not true, in general. As an example consider $X \sim N(0, 1)$ and define $X_n =$

$(-1)^n X$. Then $X_n \stackrel{D}{=} X$ and hence $X_n = O_p(1)$. But $P(|X_n| > \epsilon) = 2\Phi(\epsilon) - 1 \neq 0$ and hence we do not have the relation $X_n = o_p(1)$.

We often use representations like $X_n = Y_n + R_n$ where either $R_n \to 0$ in probability or $nR_n \to 0$ in probability. We can use small o notation to write these as $X_n = Y_n + o_p(1)$ and $X_n = Y_n + o_p(1/n)$, respectively.

4 Convergence of a series Of Random Variables

We have already discussed sequence of random variables and their convergence. However, like ordinary series, we can also define a series with random variables. The different terms of such a series is either independent or dependent. Often our interest lies in the asymptotic nature of average or some standardised version of it. Consequently, we have laws of large numbers and central limit theorems.

4.1 Weak law of large numbers & law of averages

Let $X_n, n \geq 1$ be a sequence of random variables and let $S_n = \sum_{k=1}^{n} X_k$. Consider a sequence of constants $b_n, n \geq 1$, where b_n is positive, nondecreasing and diverging to $+\infty$. Then $X_n, n \geq 1$ satisfies weak law of large numbers(WLLN) with respect to b_n, if there exists a sequence of real constants a_n such that

$$\frac{S_n - a_n}{b_n} \xrightarrow{P} 0$$

as $n \to \infty$. a_n are called centering constants and b_n are called norming constants. Most often we take $a_n = E(S_n)$ and $b_n = \sqrt{Var(S_n)}$.

Suppose a random experiment A is performed n times and f_n is the sample frequency. Naturally, nothing can be predicted about the rate of occurrence P(A) based on a single trial. However, considering the averages $\frac{f_n}{n}, n \geq 1$, we arrive at an experiment in which the outcome can be predicted with high accuracy. The Law of Large Numbers are often describes as law of averages.

4.1.1 Different WLLN

Depending on different models, different variations of WLLN exist. We provide below few of them. **Markov WLLN:** Consider a sequence of random variables X_n with $E(X_k) = \mu_k$. If $\frac{Var(S_n)}{n^2} \to 0$ then

$$\frac{S_n}{n} - \frac{\sum_{k=1}^{n} \mu_k}{n} \xrightarrow{P} 0.$$

That is, $X_n, n \geq 1$ satisfies WLLN. If, in addition, X_n are independent, then we have Chebyshev's WLLN.

Poisson's WLLN: Consider a sequence of independent random variables X_n with $X_n \sim Bin(1, \mu_n)$. Then

$$\frac{S_n}{n} - \frac{\sum_{k=1}^{n} \mu_k}{n} \xrightarrow{P} 0.$$

If $\mu_k = \mu$, then the above WLLN reduces to that of Bernoulli.

Khintchine WLLN: Consider a sequence of iid random variables X_n with $E(X_1) = \mu < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{P} \mu.$$

4.1.2 An Example

Consider tossing of a coin n times. If f_n is the number of heads turned up then $\frac{f_n}{n}$ gives the proportion of heads in n throws. Naturally, nothing is known about the probability of head at the outset. The Law of Large Numbers predicts that $\frac{f_n}{n}$ will be close to the unknown probability for large n. Thus the value of $\frac{f_n}{n}$ for large n would provide a close idea about the unknown probability. For better understanding, we simulate the tosses of a coin with success probability .7. For varying n(50,500 and 5000, clockwise), we have constructed the histogram of the proportion of heads in n trials. It is easy to observe that the distribution of the observed success proportion becomes more and more concentrated about the true proportion 0.7 as we increase n.

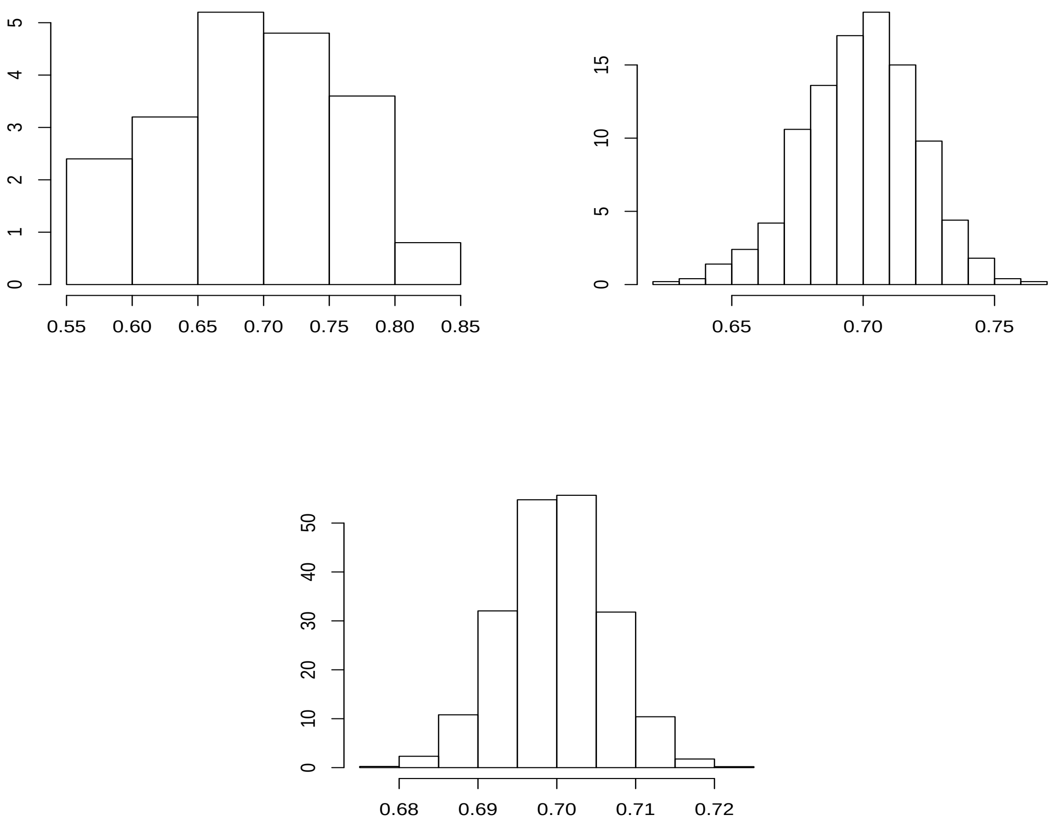


Figure 1: Convergence of observed success proportion

It is easy to observe that the distribution of the observed success proportion becomes more and more concentrated about the true proportion 0.7 as we increase n.

4.2 Strong law of large numbers(SLLN)

Let $X_n, n \geq 1$ be a sequence of random variables and let $S_n = \sum_{k=1}^{n} X_k$. Consider a sequence of constants $b_n, n \geq 1$, where b_n is positive and diverging to $+\infty$. Then $X_n, n \geq 1$ satisfies strong law of large numbers(SLLN) with respect to b_n, if there exists a sequence of real constants a_n such that

$$\frac{S_n - a_n}{b_n} \overset{a.s.}{\to} 0$$

as $n \to \infty$. Unlike WLLN, in case of SLLN, we consider $a_n = E(S_n)$ but $b_n = n$.

4.2.1 SLLN & Borel-Cantelli Lemma

Almost sure convergence is not easy to check in a straightforward way and hence we need some sufficient conditions. The following result provides an way to establish almost sure convergence.

Borel-Cantelli Lemma: If for a sequence of events $A_n, n \geq 1$, $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P(\overline{\lim} A_n) = 0$. However, if $A_n, n \geq 1$ are pairwise independent with $\sum_{n=1}^{\infty} P(A_n) = \infty$ then $P(\overline{\lim} A_n) = 1$.
Observe that

$$X_n - X \overset{a.s.}{\to} 0 \Longleftrightarrow P(\overline{\lim} |X_n - X| > \epsilon) = 0$$

for every $\epsilon > 0$ and hence we can use the above lemma to establish almost sure convergence.

Large sample Inference: Module 14[1]

What we provide in this module

- Univariate Delta Theorem with general scaling factor
- Multivariate Delta Theorems
- Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Asymptotic variance & Delta Theorem

From delta theorem, we find that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$ then under certain conditions $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, [g'(\theta)]^2\sigma^2(\theta))$. The quantity $[g'(\theta)]^2\sigma^2(\theta)$ is called the asymptotic variance(AV) of $g(T_n)$. However, the relation $AV(g(T_n)) = \lim_{n\to\infty} nVar[g(T_n)]$ may not always hold. For example, suppose X_n are iid N(1,1) variables and $T_n = \bar{X}$. Then from CLT, $\sqrt{n}(T_n - 1) \xrightarrow{D} N(0, 1)$. Define $g(T_n) = 1/T_n$, then from delta theorem $AV(1/T_n) = 1$. However, exact calculation shows $Var(1/T_n)$ does not exist and hence the proposition.

2 Scaling factor & Delta Theorems

For univariate delta theorems, we have assumed so far the scaling factor $\sqrt{n}$. But in practice, scaling factor may be other than $\sqrt{n}$. Let us explain using an example. Suppose $X_n, n \geq 1$ are iid $R(0, \theta)$ variables and let $X_{(n)}$ denote the largest order statistic. Define $Y_n = n(\theta - X_{(n)})$, then $F_{Y_n}(x) = 1-(1-\frac{x}{n\theta})^n \to 1-exp(-x/\theta)\ \forall\, x$. Hence $n(\theta - X_{(n)}) \xrightarrow{D}$ Exponential(mean $=$ θ). Thus we get a scaling factor n instead of the usual $\sqrt{n}$. Naturally, the already discussed univariate delta theorems are of no use in finding the asymptotic distribution of some function $g(X_{(n)})$.

2.1 Delta theorem of first order with general scaling factor

Result: Consider a sequence of random variables $X_n, n \geq 1$ such that $a_n(X_n - \theta) \xrightarrow{D} X$ for some $a_n > 0$ such that $a_n \to \infty$. Let $g(\theta)$ be a real valued function such that $g'(\theta) = \frac{d}{d\theta}g(\theta)$ is continuous and non zero in a neighbourhood of θ. Then

$$a_n(g(X_n) - g(\theta)) \xrightarrow{D} g'(\theta)X.$$

Proof: As earlier, $a_n(X_n - \theta) \xrightarrow{D} X$ implies $X_n - \theta \xrightarrow{P} 0$. Then using Taylor expansion, we have the representation $a_n(g(X_n) - g(\theta)) = a_n(X_n - \theta)(g'(\theta) + o_P(1))$. Since, $a_n(X_n - \theta) \xrightarrow{D} X$,

we have $a_n(g(X_n) - g(\theta)) = a_n(X_n - \theta)g'(\theta) + O_P(1)o_P(1) = a_n(X_n - \theta)g'(\theta) + o_P(1)$. Now $o_P(1) \xrightarrow{P} 0$ and hence applying Slutsky theorem , the result follows.

2.2 Delta theorem of higher order with general scaling factor

Result: Consider a sequence of random variables $X_n, n \geq 1$ such that

$$a_n(X_n - \theta) \xrightarrow{D} X$$

as $n \to \infty$. Let $g(\theta)$ be a real valued function such that $g^{(r)}(\theta) = \frac{d^r}{d\theta^r}g(\theta) = 0 \ \forall \ r = 1, 2, .., m-1$ but $g^{(m)}(\theta) \neq 0$. Then

$$m! {a_n}^m(g(X_n) - g(\theta)) \xrightarrow{D} g^{(m)}(\theta)X^m.$$

Proof: The proof of the above is an easy consequence of Taylor's theorem of higher orders with random variables and hence omitted.

2.3 An application

Suppose $X_n, n \geq 1$ are iid as $R(0, \theta), \theta > 0$ and we are interested in $g(X_{(n)}) = logX_{(n)}$. It is already known that $n(\theta - X_{(n)}) \xrightarrow{D}$ Exponential(mean=θ). Define $g(x) = log(x)$ then $g'(x) = 1/x \neq 0$. Then applying the above we get that

$$n(log(\theta) - log(X_{(n)})) \xrightarrow{D} \frac{X}{\theta},$$

where $X \sim$ Exponential(mean=θ). Since $\frac{X}{\theta}$ has a standard exponential distribution, we have that the asymptotic distribution of $n(log(\theta) - log(X_{(n)}))$ is standard exponential.

3 Delta Theorem: Multivariate case

The Theorem: Consider a sequence of p component random vectors $\mathbf{X}_n, n \geq 1$ and a fixed p component vector θ such that

$$a_n(\mathbf{X}_n - \theta) \xrightarrow{D} \mathbf{X}$$

as $n \to \infty$ for some p component random vector $\mathbf{X}$. Let $g(\theta)$ be a real valued function such that the gradient vector $\nabla g = (\frac{\partial g(\theta)}{\partial \theta_1}, \frac{\partial g(\theta)}{\partial \theta_2}, ..., \frac{\partial g(\theta)}{\partial \theta_p})^T$ is continuous and non-zero in some neighbourhood(nbd) of θ. Then

$$a_n(g(X_n) - g(\theta)) \xrightarrow{D} (\nabla g)^T \mathbf{X}.$$

Proof: Suppose that $|\mathbf{X}_n - \theta|$ denotes the Euclidean distance. Assume that ∇g is continuous in the neighbourhood(nbd) $|\mathbf{x} - \theta| < \epsilon$. Then for $|\mathbf{x} - \theta| < \epsilon$ using Mean Value Theorem(MVT), we have

$$g(\mathbf{x}) - g(\theta) = (\mathbf{x} - \theta) \int_0^1 g'(\theta + t(\mathbf{x} - \theta))dt.$$

As a consequence of the fact that $a_n(\mathbf{X}_n - \theta) \xrightarrow{D} \mathbf{X}$, we have for every $\delta > 0$, $P(|\mathbf{X}_n - \theta| < \delta) = P(a_n|\mathbf{X}_n - \theta| < \delta a_n) \to 1$ and hence $\mathbf{X}_n - \theta \xrightarrow{P} 0$.
Now for $|\mathbf{x} - \theta| < \epsilon$, we have the representation $a_n(g(\mathbf{X}_n) - g(\theta)) = a_n(\mathbf{X}_n - \theta) \int_0^1 g'(\theta + t(\mathbf{X}_n - \theta))dt$, where due to continuity and the fact that $\mathbf{X}_n \xrightarrow{P} \theta$, we have $\int_0^1 g'(\theta + t(\mathbf{X}_n - \theta)dt \xrightarrow{P} g'(\theta)$. Thus we have for vector valued random variable

$$a_n(g(\mathbf{X}_n) - g(\theta)) \xrightarrow{D} \{\nabla g\}^T \mathbf{X}.$$

In most of the cases, $\mathbf{X} \sim N_p(\mathbf{0}, \Sigma)$. Thus from the properties of linear functions of multivariate normal variables, we have the large sample distribution of $a_n(g(\mathbf{X}_n) - g(\theta))$ as $N(0, \{\nabla g\}^T \Sigma \{\nabla g\})$.

3.1 Examples

Example 1: Suppose $\sqrt{n}(T_{1n} - \theta_1, T_{2n} - \theta_2)^T \xrightarrow{D} N_2(0, \Sigma)$. We are interested in the asymptotic distribution of $g(T_1, T_2) = T_1 T_2$. Assume that $g(x, y)$ is a function of the real variables x and y, is continuous and has continuous partial derivatives in the nbd of (θ_1, θ_2). Then by the delta theorem for multiple random variables, the asymptotic distribution of

$T_{1n}T_{2n}$ is normal with mean $\theta_1\theta_2$. Now $\nabla g(\theta_1,\theta_2) = (\theta_2,\theta_1)^T$. Hence the asymptotic variance comes out as $(\theta_2^2\sigma_{11} + 2\theta_1\theta_2\sigma_{12} + \theta_1^2\sigma_{22})/n$, where $\Sigma = ((\sigma_{ij}))$.

Example 2: Suppose $\sqrt{n}(T_{1n}-\theta_1, T_{2n}-\theta_2)^T \xrightarrow{D} N_2(0,\Sigma)$ and we are interested in the asymptotic distribution of $g(T_1,T_2) = T_1/T_2$. Assume that $g(x,y)$ is a function of the real variables x and y, is continuous and has continuous partial derivatives in the nbd of (θ_1,θ_2) except for $y = 0$. Thus if (θ_1,θ_2) is such that $\theta_2 \neq 0$, then the asymptotic distribution of T_{1n}/T_{2n} is normal with mean θ_1/θ_2. However, $\nabla g(\theta_1,\theta_2) = (1/\theta_2, -\theta_1/\theta_2^2)^T$. Consequently the asymptotic variance comes out as $(\theta_1/\theta_2)^2\{\sigma_{11}/\theta_1^2 - 2\sigma_{12}/(\theta_1\theta_2) + \sigma_{22}/\theta_2^2\}/n$, provided $\theta_1 \neq 0$ and $\Sigma = ((\sigma_{ij}))$.

Example 3: Distribution of sample variance Suppose $X_i, i \geq 1$ are iid with mean μ, variance σ^2 and finite fourth order moments. Without any loss of generality, assume $\mu = 0$ and $\sigma = 1$. Note that the sample variance $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - (\bar{X}-\mu)^2$. Thus we start from the joint asymptotic distribution of $T_{1n} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)$ and $T_{2n} = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$. By multivariate CLT, we have $\sqrt{n}(T_{1n}-\theta_1, T_{2n}-\theta_2)^T \xrightarrow{D} N_2(0,\Sigma)$, where $\theta_1 = 0$, $\theta_2 = \sigma^2$ and

$$\Sigma = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

Now define $g(x,y) = y - x^2$. Then $g(x,y)$ is continuous and has continuous partial derivatives in the nbd of (θ_1,θ_2). Since $\nabla g(\theta_1,\theta_2) = (0,1)^T \neq \mathbf{0}$, we find from the delta theorem with multiple variables that that $\sqrt{n}(s^2 - \mu_2) \xrightarrow{D} N(0, (\nabla g(\theta_1,\theta_2))^T \Sigma \nabla g(\theta_1,\theta_2))$. Now after a little manipulation we find $(\nabla g(\theta_1,\theta_2))^T\Sigma\nabla g(\theta_1,\theta_2) = \mu_4 - \mu_2^2$.
If the underlying distribution is normal, then $\mu_4 = 3\sigma^4$ and hence in such a case asymptotic variance reduces to $2\sigma^4$.
As an easy consequence of the delta theorem, we find that $\sqrt{n}(s-\sigma) \xrightarrow{D} N(0, \frac{\mu_4-\mu_2^2}{4\mu_2})$, which under normality reduces to $\sqrt{n}(s-\sigma) \xrightarrow{D} N(0, \frac{\sigma^2}{2})$.

Large sample Inference: Module 12[1]

What we provide in this module

- Complete Convergence
- Different SLLN & applications
- Continuity Theorem & applications
- Central limit theorems

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Complete convergence & Almost sure convergence

Now we discuss another type of convergence, namely, complete convergence. Consider the random variables X_n and X defined on a common sample space. Then X_n is said to converge completely to X if $\sum_{n=1}^{\infty} P(d(X_n, X) > \epsilon) < \infty$ for every $\epsilon > 0$. In notation, we use $X_n - X \xrightarrow{c} 0$.

Complete convergence is useful in establishing almost sure convergence. First of all note that $X_n - X \xrightarrow{c} 0$ implies $X_n - X \xrightarrow{a.s.} 0$ but the converse may not hold. Thus from the properties of convergent series, if one can show that $P(d(X_n, X) > \epsilon) = O(n^{1+r})$ for some $r > 0$, then complete convergence and hence almost sure convergence holds.

2 Different variations of SLLN

Depending on different models, we have different variations of WLLN and we provide below few of them.

2.1 Borel SLLN

Result: Let $X_n, n \geq 1$ be iid $Bernoulli(\theta)$ variables. Then $\frac{S_n}{n} \xrightarrow{a.s.} \theta$.

Proof: Note that $S_n \sim Bin(n, \theta)$ and using the properties of binomial distribution, one can find that $E(\frac{S_n}{n} - \theta)^2 = \theta(1-\theta)/n$. Then by Markov's inequality $P(|\frac{S_n}{n} - \theta| > \epsilon) \leq E(\frac{S_n}{n} - \theta)^2/\epsilon^2$ and since $E(\frac{S_n}{n} - \theta)^2 \leq \frac{1}{4n}$, nothing can be concluded.

However, if we consider $E(\frac{S_n}{n} - \theta)^4 = n^{-3}\theta(1-\theta)[1 + 3\theta(1-\theta)(n-2)]$. Since $\theta(1-\theta) \leq 1/4$, we get $1 + 3\theta(1-\theta)(n-2) \leq (3n-2)/4 < 3n/4$ and hence we get $E(\frac{S_n}{n} - \theta)^4 \leq 3n^{-2}/16$, that is, $E(\frac{S_n}{n} - \theta)^4 = O(n^2)$. Thus $P(|\frac{S_n}{n} - \theta| > \epsilon) \leq E(\frac{S_n}{n} - \theta)^4/\epsilon^4 \leq 3n^{-2}/16\epsilon^4$ for any $\epsilon > 0$. Since the series $\sum_{n=1}^{\infty} P(d(X_n, X) > \epsilon)$ converges, the result follows.

2.2 Other SLLN

Kolmogorov SLLN: Let $X_n, n \geq 1$ be independent random variables with finite variances. Then $\frac{S_n}{n} - \frac{E(S_n)}{n} \stackrel{a.s.}{\to} 0$ if $\sum_{n=1}^{\infty} \frac{Var(X_n)}{n^2} < \infty$.

SLLN for bounded random variables: If $X_n, n \geq 1$ be iid uniformly bounded random variables then $\frac{S_n}{n} \stackrel{a.s.}{\to} E(X_1)$.

Khinthine SLLN: Let $X_n, n \geq 1$ be iid random variables with finite expectation. Then $\frac{S_n}{n} \stackrel{a.s.}{\to} E(X_1)$.

Now we consider an extension of the above in the context of multiple random variables. Let $X_n, n \geq 1$ be iid q component random variables with finite expectation θ. If g is continuous at θ then $g(\bar{X}_n) \stackrel{a.s.}{\to} g(\theta)$.

An application will be helpful to understand the idea. Let $X_n, n \geq 1$ be iid random variables with finite variance σ^2 and expectation μ. Consider $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Now $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\}^2$. Define $\mathbf{Z} = (Z_{1i}, Z_{2i})$ with $Z_{1i} = (X_i - \mu)^2$ and $Z_{2i} = (X_i - \mu)$. Then $\bar{\mathbf{Z}} \stackrel{a.s.}{\to} (\sigma^2, 0)'$. Since $g(x, y) = x - y^2$ is continuous, we have the desired result from the above.

3 Continuity theorem

Let $X_n, n \geq 1$ be a sequence of random variables with $F_n, n \geq 1$ as the corresponding sequence of cdf's. If $M_n(t), n \geq$ is the corresponding sequence of MGFs then the following questions are natural:

Does $M_n(t)$ converge to an MGF for large n?

If $X_n \stackrel{D}{\to} X$, for some rv X with MGF M(t), does $M_n(t) \to M(t)$!

Continuity theorem, given below, provides the answers to the above.

The theorem: Let $F_n, n \geq 1$ be a sequence of DFs with $M_n(t), n \geq$ as the corresponding sequence of MGFs. If there exists a DF F corresponding to MGF M such that $M_n(t) \to M(t)$

as $n \to \infty$ then $F_n \xrightarrow{w} F$.

3.1 Applications of Continuity theorem

Example 1: Let X_n be such that $P(X_n = 1) = n^{-2} = 1 - P(X_n = 2)$. Then $M_n(t) = E(e^{tX_n}) = \frac{e^t}{n^2} + e^{2t}(1 - 1/n^2) \to e^{2t}$. Clearly the limit is the MGF of a random variable having degeneracy at 2. Thus we get from the above that $X_n \xrightarrow{D} X$ with $P(X = 2) = 1$.

Example 2: Let X_n be iid Bin(1,p) variables. Then $S_n = \sum_{k=1}^{n} X_k$ has the MGF $M_n(t) = (1 - p + pe^t)^n$. If we increase n in such a way that $np = \lambda$ is finite then $(1 - p + pe^t)^n \to e^{\lambda(e^t - 1)}$, which is the MGF of a Poisson random variable with mean λ. Hence $S_n \xrightarrow{D}$ Poisson(λ).

4 Central limit theorems

WLLN and SLLN provide an idea about the limit of an average of random variables. But it does not provide the rate at which the limit is reached. Central limit theorems enable us to get an idea about the rate of convergence to the limiting quantity. Now we shall discuss convergence of $\frac{S_n - a_n}{b_n}$ to a nondegenerate random variable for suitable choices of centering and norming constants. We also investigate the properties of such a nondegenerate limit.
The earliest form of CLT was postulated by Abraham de Moivre(1733) who used the normal approximation to the distribution of the number of heads of obtained through a large number of throws of an unbiased coin. However, the result of de Moivre was referenced by Pierre-Simon Laplace in his 1812 publication Thorie analytique des probabilits. Laplace extended the idea of De Moivre to approximate a binomial distribution by the normal distribution. However, the modern form of CLT is due to Aleksandr Lyapunov(1901), who developed the general form together with proof. The particular term "central limit theorem" is due to George Plya(1920) because of its central role in probability theory. However, Le Cam interpreted the word "central" because "it describes the behaviour of the centre of the

distribution as opposed to its tails"(Lecam,1986). Now we discuss different forms of CLT.

4.1 De Moivre's Limit Theorem

Let X_n be iid sequence of $Binomial(1, \mu)$ random variables. Then as $n \to \infty$, $\bar{X}_n \xrightarrow{P} \mu$ and

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \mu(1-\mu))).$$

For a proof one can expand the concerned MGF in a Taylor series and take the limit to get the MGF of the normal variable as the limit. Then the result is immediate as a consequence of continuity theorem. This is the primitive version of CLT (De Moivre,1733, Laplace, 1810), which gives the asymptotic distribution only for a sequence of iid Bernoulli variables.

4.2 Lindeberg-Levy CLT

Let X_n be iid sequence of random variables with mean μ and finite variance σ^2. Then as $n \to \infty$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Thus the limiting distribution of a normalized average is normal provided the observations are iid with finite variance. Thus the limiting distribution is independent of the parent population and hence enables to develop asymptotic procedures for a large class of parents only with finite variance assumption.

4.3 Lyapounov condition & CLT

Lindeberg-Levy CLT considers the sum of iid random variables with finite variance. But random variables may be independent but not distributed identically. In such a case we have Lyapounov CLT which assumes existence of moments higher than order 2.

Let X_n be independent sequence of random variables with mean $E(X_n) = \mu_n$ and

$Var(X_n) = \sigma_n^2$. If for some $\delta > 0$, the Lyapounov condition

$$(\sum_{k=1}^{n} \sigma_k^2)^{-2-\delta} \sum_{k=1}^{n} E|X_k - \mu_k|^{2+\delta} \to 0$$

as $n \to \infty$ is satisfied then with $s_n^2 = \sum_{k=1}^{n} \sigma_k^2$,

$$\sum_{k=1}^{n} \frac{X_k - \mu_k}{s_n} \xrightarrow{D} N(0, 1).$$

4.4 Lindeberg condition & CLT

Lindebergs condition is an alternative to Lyapounovs condition for satisfying CLT. If Lyapounovs condition is satisfied then Lindeberg condition is also satisfied but the converse is not always true.

Let X_n be independent sequence of random variables with mean $E(X_n) = \mu_n$ and $Var(X_n) = \sigma_n^2$. If for every $\epsilon > 0$, the Lindeberg condition

$$(\sum_{k=1}^{n} \sigma_k^2)^{-2} \sum_{k=1}^{n} E(X_k - \mu_k)^2 I(|X_k - \mu_k| > \epsilon \sqrt{\sum_{k=1}^{n} \sigma_k^2}) \to 0$$

as $n \to \infty$ is satisfied then

$$\sum_{k=1}^{n} \frac{X_k - \mu_k}{s_n} \xrightarrow{D} N(0, 1).$$

5 Uniform convergence in CLT

CLT only provides the limiting distribution but not the choice of n and accuracy of approximation. However, choice of an n ensuring limit distribution to be normal depends on the underlying distribution. Suppose $G_n(x)$ is the cdf of $\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma}$ and $\Phi(x)$ is the DF of N(0,1) variable. Then CLT states that $G_n(x) \to \Phi(x)$ for all x but the convergence is not uniform so that for fixed n, the approximation may perform poorly. The following result gives a bound on the margin of error $|G_n(x) - \Phi(x)|$ and shows that the convergence is uniform for the class of distributions with finite third order absolute moments.

Berry-Essen Bound: Let X_n be iid sequence of random variables with mean μ, variance σ^2 and finite absolute third order moment $\rho = E|X_1 - \mu|^3$. Then for all n

$$|G_n(x) - \Phi(x)| < \frac{c\rho}{\sigma^3\sqrt{n}} \quad \text{for all } x \Leftrightarrow \sup_x |G_n(x) - \Phi(x)| < \frac{c\rho}{\sigma^3\sqrt{n}},$$

where the constant $c \in (.4097, .7975)$ depends on the underlying distribution but is independent of n.

Berry-Essen theorem gives a sufficient condition for CLT even for non identical class of distributions. Thus we find a sufficient condition for CLT.

Let $X_n \sim F_n$ be an independent sequence of random variables with $E(X_n) = \mu_n$, $Var(X_n) = \sigma_n^2$ such that $\frac{E_{F_n}|X_1-\mu_n|^3}{\sigma_n^3} = o(\sqrt{n})$. Then $G_n(x) \to \Phi(x)$ as $n \to \infty$.

For an application, consider an independent sequence of random variables X_n such that $X_n \sim Bernoulli(p_n)$. Then $F_n = Bernoulli(p_n)$ and $E_{F_n}|X_1 - p_n|^3 \leq 1$ for any n. Thus the CLT holds as long as $p_n(1-p_n)$ converges to a non zero quantity. However, $\frac{E_{F_n}|X_1-\mu_n|^3}{\sigma_n^3} = \frac{p_n^2+(1-p_n)^2}{\sqrt{p_n(1-p_n)}}$ and hence if $(np_n)^{-1/2} \to 0$, the limiting distribution approaches normal.

Large sample Inference: Module 16[1]

What we provide in this module

- Asymptotic distribution of kurtosis coefficient
- Asymptotic joint distribution of skewness & kurtosis coefficient
- Asymptotic distribution of sample correlation coefficient
- Rate of convergence

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Further applications of delta theorem

Example 1: Distribution of sample kurtosis coefficient

Suppose $X_n, n \geq 1$ are iid random variables with finite eighth order moment μ_8. If we denote the r th order central moment by m_r then the population kurtosis coefficient is $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$ provided $\mu_2 > 0$. As earlier, $m_r \xrightarrow{P} \mu_r$ for any r, and hence a consistent estimate of γ_2 is $g_2 = \frac{m_4}{m_2^2} - 3$ provided $m_2 > 0$. Then g_2 is a function of two sample central moments, namely, m_2 and m_4. Now, to derive the asymptotic distribution, we define $g(u, v) = \frac{v}{u^2} - 3$, which is continuous and has continuous derivatives except for $u = 0$. Now from the joint asymptotic distribution of the sample central moments, we have $\sqrt{n}(m_2 - \mu_2, m_4 - \mu_4)^T \xrightarrow{D} N_2(\mathbf{0}, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_{22} & \sigma_{24} \\ & \sigma_{44} \end{pmatrix}$.

Now a straightforward algebra gives

$$\nabla g(\mu_2, \mu_4) = (-\frac{2\mu_4}{\mu_2^3}, \mu_2^{-2})^T$$

and hence we obtain $\sqrt{n}(g_2 - \gamma_2) \xrightarrow{D} N(\mathbf{0}, \tau_2^2)$, where

$$\tau_2^2 = (\nabla g)^T \Sigma \nabla g = (\gamma_2 + 3)^2 h(\mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7, \mu_8)$$

for some function h. We refer the interested reader to the book by Serfling(1980) for an explicit expression of h. However, if the underlying distribution is standard normal (due to location and scale invariance, one can WLG consider this distribution), then $\mu_3 = 0, \mu_5 = 0, \mu_7 = 0$, $\mu_4 = 3, \mu_6 = 15, \mu_8 = 105$ and hence we obtain $\sqrt{n}(g_2 - \gamma_2) \xrightarrow{D} N(0, 24)$.

Example 2: Joint distribution of sample skewness & kurtosis coefficients

Suppose $X_n, n \geq 1$ are iid random variables with finite eighth order moment μ_8. Then with the already used notations, the sample skewness and kurtosis coefficients are $g_1 = \frac{m_3}{m_2^{3/2}}$ and $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$, respectively, provided $\mu_2 > 0$. Now we shall obtain the joint asymptotic distribution of such coefficients. We define a vector valued function $\mathbf{g}(u_1, u_2, u_3) = \begin{pmatrix} g_1(u_1, u_2, u_3) \\ g_2(u_1, u_2, u_3) \end{pmatrix}$ where $g_1(u_1, u_2, u_3) = u_2/u_1^{3/2}$ and $g_2(u_1, u_2, u_3) = u_3/u_1^2 - 3$. As indi-

cated earlier, each g_i is continuous and has continuous derivatives except for $u_1 = 0$. Now from the joint asymptotic distribution of the sample central moments, we have $\sqrt{n}(m_2 - \mu_2, m_3 - \mu_3, m_4 - \mu_4)^T \xrightarrow{D} N_2(\mathbf{0}, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_{22} & \sigma_{23} & \sigma_{24} \\ & \sigma_{33} & \sigma_{34} \\ & & \sigma_{44} \end{pmatrix}$.

Now we define the matrix $G^{2X3} = ((\frac{\partial g_i}{\partial u_j}\Big|_{(\mu_2,\mu_3,\mu_4)}))$. A straightforward algebra gives

$$G = \begin{pmatrix} -\frac{3\mu_3}{2\mu_2^{5/2}} & \mu_2^{-3/2} & 0 \\ -\frac{2\mu_4}{\mu_2^3} & 0 & \mu_2^{-2} \end{pmatrix}.$$

Hence we obtain $\sqrt{n}(g_1 - \gamma_1, g_2 - \gamma_2)^T \xrightarrow{D} N(\mathbf{0}, \Psi)$, where $\Psi = G\Sigma G^T$. The exact expression of Ψ requires a huge calculation and hence we skip this. However, if the underlying distribution is standard normal (due to location and scale invariance, one can WLG consider this distribution), then $\mu_3 = 0, \mu_5 = 0, \mu_7 = 0$, $\mu_4 = 3, \mu_6 = 15, \mu_8 = 105$ and hence we obtain $\Sigma = \begin{pmatrix} 2 & 0 & 12 \\ & 6 & 0 \\ & & 108 \end{pmatrix}$.

Again we obtain

$$G = \begin{pmatrix} 0 & 1 & 0 \\ -6 & 0 & 1 \end{pmatrix}.$$

Thus we obtain that

$$\sqrt{n}(g_1 - \gamma_1, g_2 - \gamma_2)^T \xrightarrow{D} N(\mathbf{0}, Diag(6, 24)).$$

Thus we find that the sample skewness and kurtosis coefficients are asymptotically independent if the underlying distribution is normal. However, one can check that if the underlying distribution is symmetric about the origin with finite even order moments up to and including the order 8, independence holds.

Example 3: Distribution of sample correlation coefficients

Suppose $(X_i, Y_i), i \geq 1$ are iid random vectors with $E(X_i^4) < \infty$ and $E(Y_i^4) < \infty$. Define the correlation coefficient based on n pairs of observations by $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$, where

$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. Naturally, r_n is invariant to location transformation and hence WLG, we can assume $E(X_i) = E(Y_i) = 0\ \forall i$. For nonnegative integers r and s define $\mu_{rs} = E(X_1^r Y_1^s)$ and $m_{rs} = \frac{1}{n}\sum_{i=1}^{n} X_i^r Y_i^s$. Then $s_{xy} = m_{11} - m_{10}m_{01}$, $s_{xx} = m_{20} - m_{10}^2$ and $s_{yy} = m_{02} - m_{01}^2$. Thus we observe that r_n is a function of five moments, namely, $m_{11}, m_{20}, m_{02}, m_{01}, m_{10}$. Now it follows from CLT that $\sqrt{n}(m_{01} - \mu_{01}, m_{10} - \mu_{10}, m_{20} - \mu_{20}, m_{02} - \mu_{02}, m_{11} - \mu_{11})^T \xrightarrow{D} N_5(\mathbf{0}, \Sigma)$ with $\Sigma = Disp(X_1, Y_1, X_1^2, Y_1^2, X_1Y_1)$. Naturally, $\mu_{01} = \mu_{10} = 0$.

Define the vector valued function $\mathbf{g}(u_1, u_2, u_3, u_4, u_5) = \begin{pmatrix} u_3 - u_1^2 \\ u_4 - u_2^2 \\ u_5 - u_1u_2 \end{pmatrix}$. Thus with the already introduced notations, we have $G^{3X5} = ((\frac{\partial g_i}{\partial u_j}\Big|_{(0,0,\mu_{20},\mu_{02},\mu_{22})}))$. Now a straightforward algebra gives

$$G = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence applying delta theorem, we get $\sqrt{n}(s_{xx} - \mu_{20}, s_{yy} - \mu_{02}, s_{xy} - \mu_{11})^T \xrightarrow{D} N_3(\mathbf{0}, G\Sigma G^T)$. A simple matrix multiplication yields $G\Sigma G^T = Disp(X_1^2, Y_1^2, X_1Y_1)$. Thus we have obtained the joint asymptotic distribution of s_{xx}, s_{yy} and s_{xy}.

Finally, define the univariate function $h(u_1, u_2, u_3) = \frac{u_3}{\sqrt{u_1u_2}}$. Then a simple algebra gives

$$\nabla h(\mu_{20}, \mu_{02}, \mu_{22}) = (-\frac{\rho}{2\mu_{20}}, -\frac{\rho}{2\mu_{02}}, \frac{1}{\sqrt{\mu_{20}\mu_{02}}})^T,$$

where, $\rho = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}$. Thus by delta theorem, we get $\sqrt{n}(r_n - \rho) \xrightarrow{D} N(0, \sigma^2)$, where $\sigma^2 = (\nabla h(\mu_{20}, \mu_{02}, \mu_{22}))^T G\Sigma G^T (\nabla h(\mu_{20}, \mu_{02}, \mu_{22}))$. To obtain the exact expression of σ^2, we need to perform the above matrix multiplication and hence we omit the details.

Naturally, the large sample variance depends on the unknown moments of the underlying distribution. However, the expression for the underlying distribution as bivariate normal is of special interest. Thus we assume that $(X_1, Y_1) \sim N_2(0, 0, 1, 1, \rho)$. First of all we derive the

expression of $Disp(X_1^2, Y_1^2, X_1Y_1)$ under bivariate normal assumption. It is easy to observe that $Var(X_1^2) = Var(Y_1^2) = 2$. Now $Cov(X_1^2, Y_1^2) = E(X_1^2Y_1^2) - 1$. Applying the fact that $Y_1|X_1 \sim N(\rho X_1, 1-\rho^2)$, we obtain $E(X_1^2Y_1^2) = E\{X_1^2(1-\rho^2+\rho^2X_1^2)\} = 1+2\rho^2$ and consequently $Cov(X_1^2, Y_1^2) = 2\rho^2$. Applying similar arguments and the fact the distribution of (X_1, Y_1) is exchangeable, one can obtain $Cov(X_1^2, X_1Y_1) = 2\rho = Cov(Y_1^2, X_1Y_1)$. Finally $Var(X_1Y_1) = E(X_1^2Y_1^2) - \rho^2 = 1+\rho^2$.
Thus we get

$$Disp(X_1^2, Y_1^2, X_1Y_1) = \begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ & 2 & 2\rho \\ & & 1+\rho^2 \end{pmatrix}.$$

Again for a bivariate normal parent we get

$$\nabla h(\mu_{20}, \mu_{02}, \mu_{22}) = (-\frac{\rho}{2}, -\frac{\rho}{2}, 1)^T.$$

A simple matrix multiplication finally gives $\sqrt{n}(r_n - \rho) \xrightarrow{D} N(0, (1-\rho)^2)$.

After getting the asymptotic distribution, it is natural to study the rate of convergence. In other words we need an idea about the value of the sample size ensuring asymptotic normality. For numerical evaluation ,we assume a bivariate normal parent with correlation coefficient $\rho = 0$. For $\rho = 0$, it is well known that $r\sqrt{n-2}/\sqrt{1-r^2}$ has a t distribution with df n-2. Thus it is natural to compare the asymptotic distribution with the actual. Hence we provide the normal QQ plot of the variables $T_{1n} = \sqrt{n}(r-\rho)/(1-\rho^2)$ and $T_{2n} = r\sqrt{n-2}/\sqrt{1-r^2}$ for $n = 15, 20$ based on a simulation study. The plots can be found in the next page.

Looking at the figures, we find that asymptotic normality holds even for n=15 or 20. Thus inferential procedures for n=15 or 20 observations can be based on this approximation and one can assure that the loss will be a very little. However, it is interesting to observe that the second variable whose exact distribution is t is also very close to a standard normal for

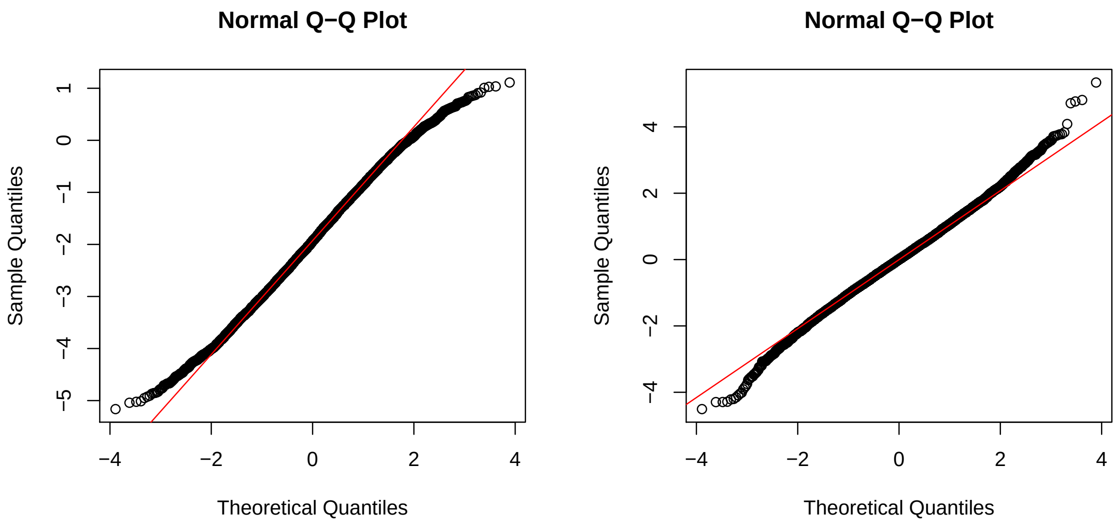


Figure 1: **QQ plot with n=15 for T_1 and T_2(left to right)**

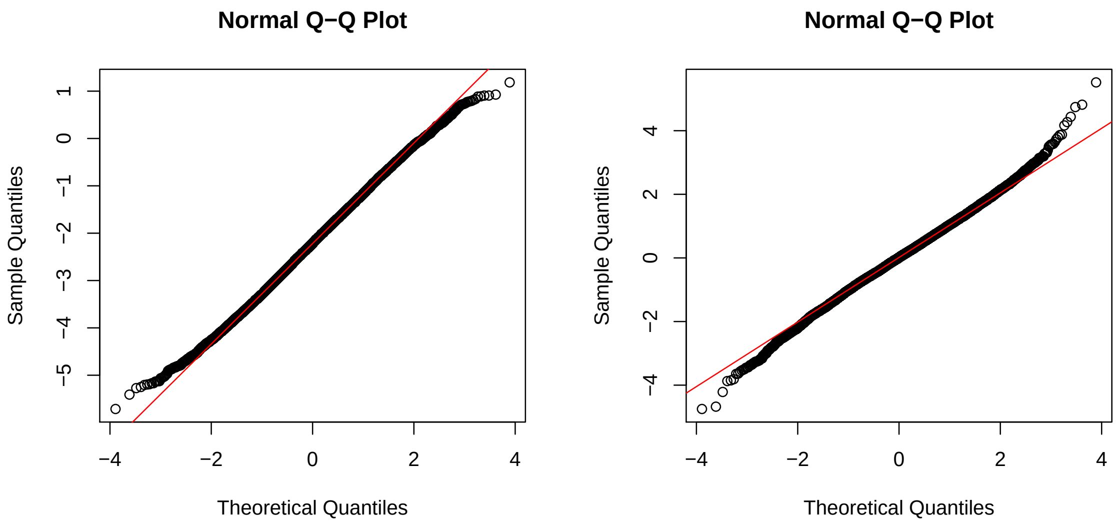


Figure 2: **QQ plot with n=20 for T_1 and T_2(left to right)**

those sample sizes. But the results obtained for $\rho = 0$ can not be stated in a generalised way.

After getting the asymptotic distribution, it is natural to study the rate of convergence. In other words we need an idea about the value of the sample size ensuring asymptotic normality. For numerical evaluation ,we assume a bivariate normal parent with correlation coefficient $\rho = 0.5$ and generate $T_{1n} = \sqrt{n}(r - \rho)/(1 - \rho^2)$ for different values of n and construct the normal QQ plot. The plots can be found in the next page.

Looking at the nature of the QQ plots, we find that as we increase n, the convergence becomes better. That is the limiting distribution becomes closer to the standard normal distribution. For n=120, we find that the approximation is good enough. However, we have used the theoretical value of ρ and hence the asymptotic variance to generate the QQ plot of T_{1n}. But in practice, asymptotic variance is not known and hence to be replaced by suitable estimates. Such replacement, of course, makes the convergence rate slower.

We have just discussed the convergence to normality for bivariate normal parent. But distributions can be non normal also. So, we consider a particular type of bivariate distribution, namely, Mckay Gamma, defined by the density

$$f(x, y) = \frac{c^{-a-b}}{\Gamma a \Gamma b} x^{a-1} e^{-y/c} (y - x)^{b-1}, \ 0 < x < y < \infty.$$

For such a distribution, the correlation coefficient is simply $\sqrt{a/(a+b)} > 0$. Thus to study the convergence rate for non normal parent, we perform a simulation study with $a = 1, b = 3$ and $c = 2$ for different values of n. In particular, we prepare histograms of $\sqrt{n}(r - .5)$ for different choices of n based on the samples from the bivariate gamma distribution. All these can be found in the next page.

We have chosen parameters in such a way that $\rho = .5$ is satisfied. We have already studied such a behaviour for bivariate normal parent. However, for bivariate gamma parent, the asymptotic variance calculation requires knowledge of the higher order moments and

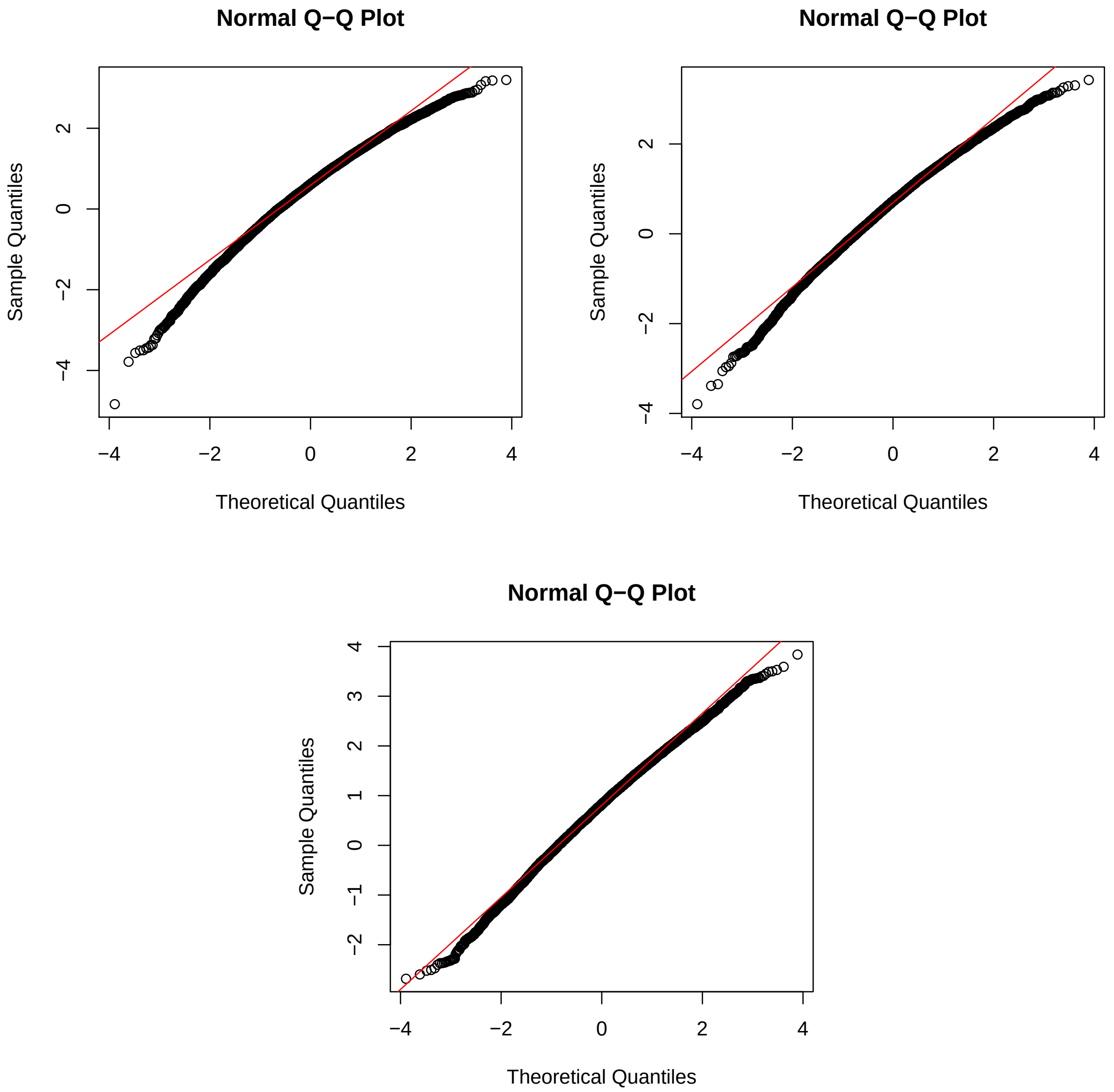


Figure 3: **Convergence to normality with $\rho = .5$ for n=60,90,120(clockwise)**

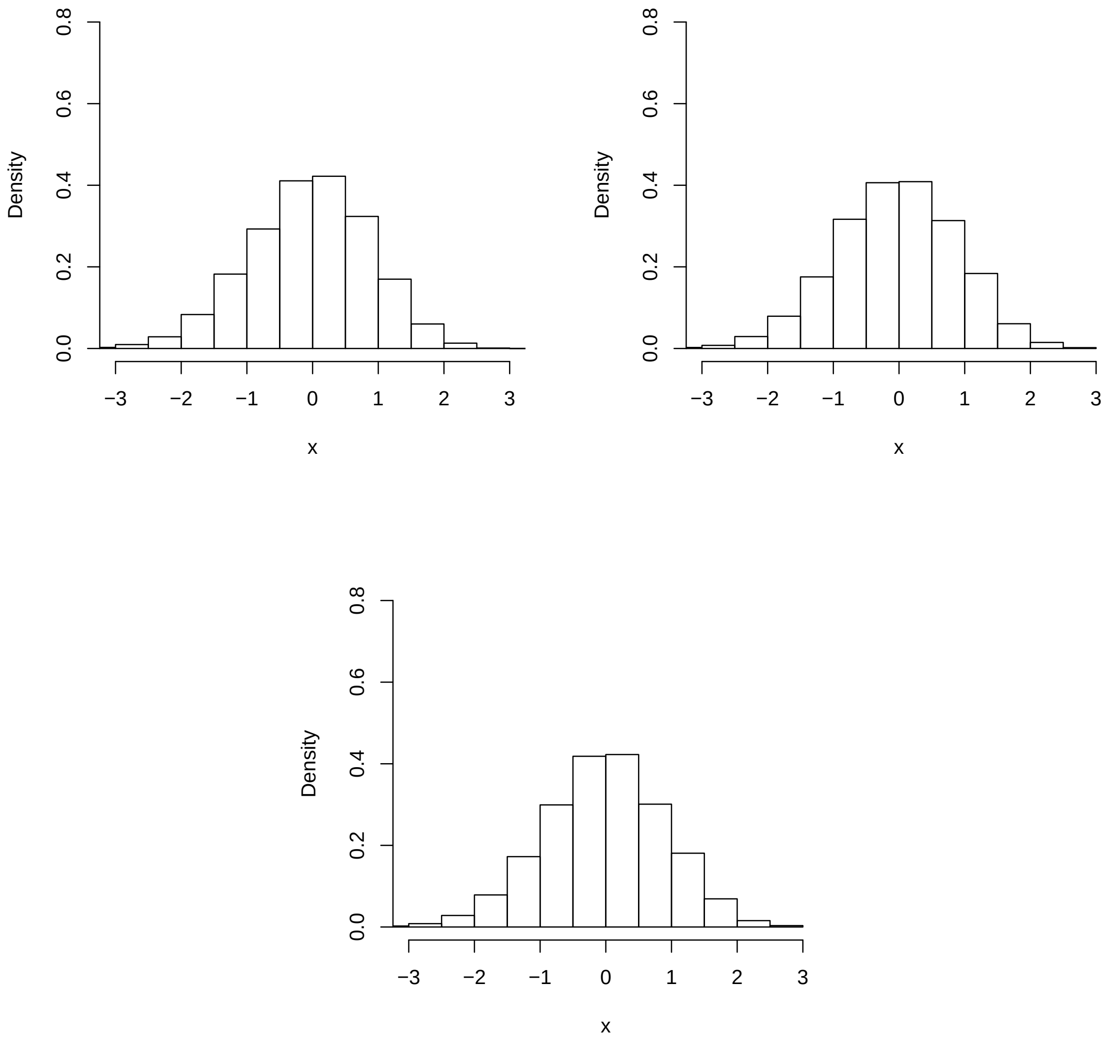


Figure 4: **Convergence to normality for n=100,150,200(clockwise)**

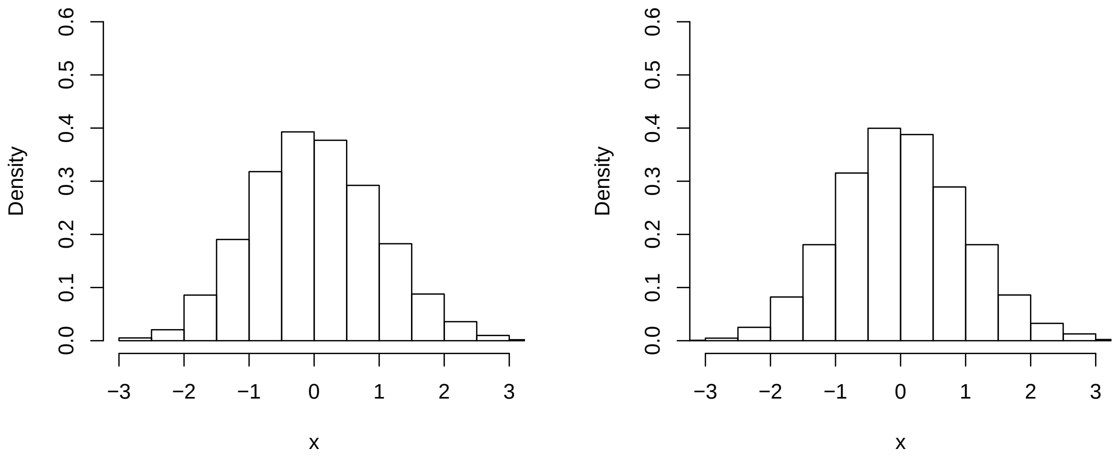


Figure 5: **Convergence to normality for** $(p_1 = .1, p_2 = .2)$**n=40,50(left to right)**

hence, for brevity, we provided only the histograms. The histogram for n=100 shows a considerable amount of skewness. However, the extent of skewness diminishes as we increase n from 100 to 200. For n=200, we arrive at the histogram of a symmetric distribution. Such a phenomena was observed for bivariate normal parent with smaller sizes. Thus, one can conclude that, non normal parent, makes the rate of convergence slower.

We have discussed so far convergence considering a continuous bivariate parent. But the underlying distribution can be discrete as well. So to know the rate of convergence for discrete parents, we consider $Multinomial(n, p_1, p_2)$ distribution and prepare the histogram of the quantity $\sqrt{n}(r + \sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}})$ for known values of p_1, p_2 and increasing values of n. The values of p_1, p_2 are chosen in a way to reflect high, moderate and low correlation coefficients. For each value of of (p_1, p_2) we simulate the value of the above quantity taking n as 40 and 80. All these can be found in the next few slides.

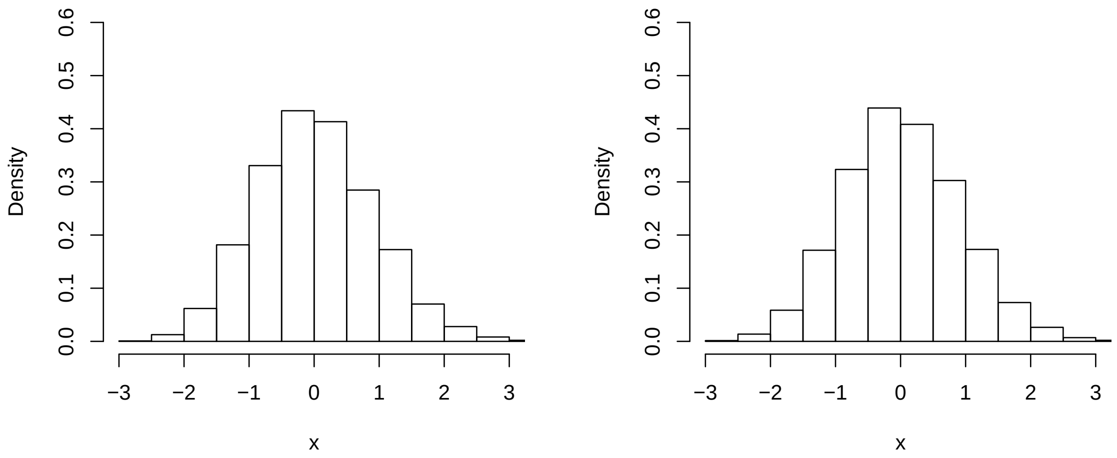


Figure 6: **Convergence to normality for** $(p_1 = .2, p_2 = .3)$**n=40,50(left to right)**

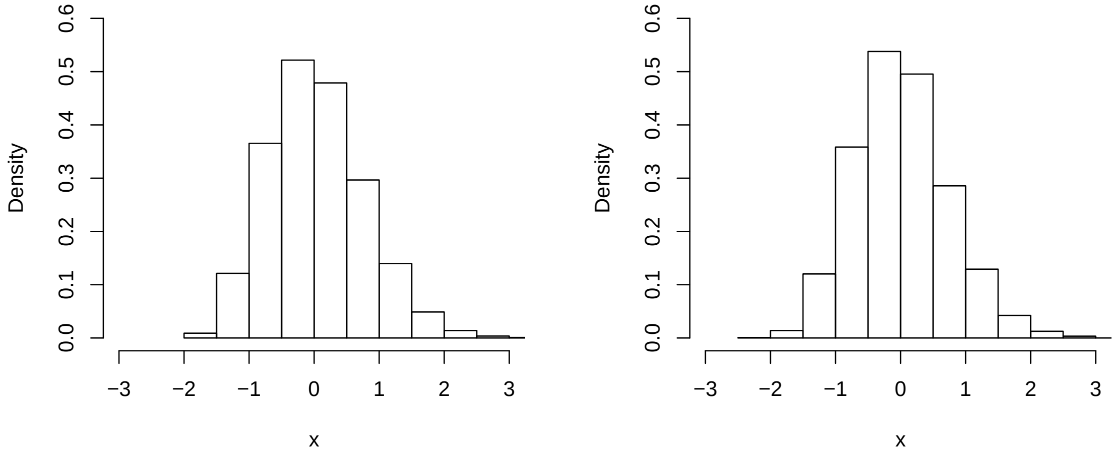


Figure 7: **Convergence to normality for** $(p_1 = .3, p_2 = .4)$**n=40,50(left to right)**

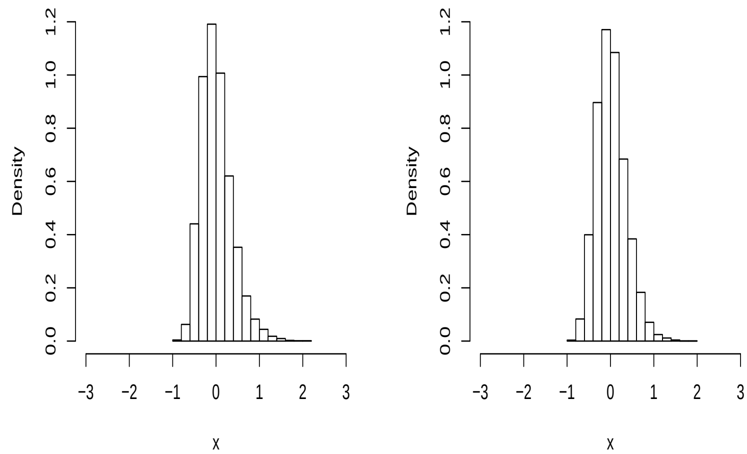


Figure 8: **Convergence to normality for** $(p_1 = .4, p_2 = .5)$**,n=40,50(left to right)**

We have fixed parameters in such a way that ρ takes the values -.17, -.33,-.53 and -.82. As expected, with the increase of n, the histogram closely represent a symmetric distribution. However, the height of the histogram increases with the increase in the theoretical correlation coefficient. Naturally, the spread reduces with such an increase. Thus, it is easy to say that the variability of the asymptotic distribution decreases with increase in the correlation coefficient value. Thus, for discrete bivariate parent, the variability of the asymptotic distribution is expected to depend on the theoretical correlation coefficient.

Large sample Inference: Module 17[1]

What we provide in this module

- Asymptotic distribution of sample quantiles
- Asymptotic distributions of quantile based measures
- Asymptotic distributions of extreme sample quantile

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Asymptotic distribution of sample quantiles

Let $X_i, i = 1, 2, .., n$ be iid observations from a continuous cdf F(x). Then all the observations are distinct with probability one. Suppose $X_{(1)} < X_{(2)} < .. < X_{(n)}$ denotes the full set of order statistics and in particular $X_{(r)}$ is the rth order statistic. For $p \in (0, 1)$, the p th order population quantile is defined as $\xi_p = F^{-1}(p)$ and the corresponding sample quantile is defined as $X_{(k)}$, where $k = [np + 1]$. We shall obtain the joint asymptotic distribution of $X_{(r)}$ and $X_{(s)}$. Note that order statistics are not explicit functions and hence delta theorem is not straightway applicable.

It is well known that if $X \sim F$ and F is continuous then $F^{-1}(U) \stackrel{D}{=} X$ for $U \sim R(0, 1)$. Thus for a full set of n observations, we find $U_{(j)} = F(X_{(j)}), j = 1, 2, ..n$, where $U_{(j)}, j = 1, 2, ..n$ is the full set of order statistics from R(0,1) population. So we first find the asymptotic joint distribution of order statistics from a R(0,1) population and then use delta methods to obtain the joint asymptotic distribution of the general order statistics. Now we provide few auxiliary results from distribution theory.

Lemma 1: If $Y_i \stackrel{iid}{\sim} Exp(mean = 1), i = 1, 2, .., n + 1$, then

$$(\frac{Y_1}{\sum_{i=1}^{n+1} Y_i}, \frac{Y_1 + Y_2}{\sum_{i=1}^{n+1}}, .., \frac{Y_1 + Y_2 + .. + Y_n}{\sum_{i=1}^{n+1} Y_i}) \stackrel{D}{=} (U_{(1)}, U_{(2)}, .., U_{(n)})$$

where $U_{(j)}, j = 1, 2, ..n$ is the full set of order statistics from R(0,1) population.

Proof: First of all, note that the joint pdf of$Y_i, i = 1, 2, .., n + 1$ is

$$f(y_1, y_2, .., y_{n+1}) = exp(-\sum_{i=1}^{n+1} y_i)I(0 < y_i < \infty, i = 1, 2, ..n).$$

Let us transform $Z_j = Y_1 + Y_2 + .. + Y_j, j = 1, 2, .., n + 1$ so that we get

$$\begin{aligned} Y_1 &= Z_1 \\ Y_j &= Z_j - Z_{j-1}, j = 2, 3, .., n + 1. \end{aligned}$$

Then we naturally get the Jacobian matrix

$$J(\frac{y}{z}) = \begin{pmatrix} 1 & -1 & 0 & .. & 0 \\ 0 & 1 & -1 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & .. & 1 & -1 \end{pmatrix}$$

Now a simple expansion through the co-factors of the first row in a recursive way gives $|J(\frac{y}{z})| = 1$. Since $Y_i > 0 \forall i$, we have $0 < Z_1 < Z_2 < .. < Z_{n+1} < \infty$ and hence we get the joint distribution of $(Z_1, Z_2, .., Z_{n+1})$ as

$$f(z_1, z_2, .., z_{n+1}) = exp(-z_{n+1}) I(0 < z_1 < z_2 < .. < z_{n+1} < \infty).$$

Next we apply the transformation

$$\begin{aligned} V_j &= \frac{Z_j}{Z_{n+1}}, j = 1, 2, .., n \\ V_{n+1} &= Z_{n+1}, \end{aligned}$$

so that we get $Z_j = V_j V_{n+1}, j = 1, 2, .., n$ and $Z_{n+1} = V_{n+1}$. Thus we naturally get the Jacobian matrix

$$J(\frac{z}{v}) = \begin{pmatrix} v_{n+1} & 0 & 0 & .. & v_1 \\ 0 & v_{n+1} & 0 & .. & v_2 \\ .. & .. & .. & .. & .. \\ 0 & 0 & .. & v_{n+1} & v_n \\ 0 & 0 & .. & & 1 \end{pmatrix}.$$

Now expanding through the co-factors of the last column, we get $|J(\frac{z}{v})| = v_{n+1}^n$. Hence we get the joint distribution of $V_j, j = 1, 2, .., n+1$ as

$$f(v_1, v_2, .., v_{n+1}) = n! I(0 < v_1 < v_2 < .. < v_n < 1) \frac{exp(-v_{n+1}) v_{n+1}^n}{\Gamma(n+1)} I(v_{n+1} > 0).$$

Hence $(V_1, ..., V_n)$ is independent of V_{n+1}, where the joint distribution of $(V_1, ..., V_n)$ is simply that of n order statistics from R(0,1) population. However, $V_{n+1} \sim Gamma(rate = 1, shape = n+1)$ and this completes the proof.

Observation: From the above result, it is evident that

$$(\frac{\sum_{i=1}^{r} Y_i}{\sum_{i=1}^{n+1} Y_i}, \frac{\sum_{i=1}^{s} Y_i}{\sum_{i=1}^{n+1} Y_i}) \stackrel{D}{=} (U_{(r)}, U_{(s)})$$

for $r < s$. Thus we have expressed the equivalence of joint distribution of order statistics with a function of sums of iid random variables. This equivalence enables us to apply CLT on the equivalent sets of random variables.

Lemma 2: If $\sqrt{n}(r/n - p_1) \to 0$ and $\sqrt{n}(s/n - p_2) \to 0$ with $0 < p_1 < p_2 < 1$ then

$$\sqrt{n} \begin{pmatrix} \frac{T_r}{n+1} - p_1 \\ \frac{T_s - T_r}{n+1} - (p_2 - p_1) \\ \frac{T_{n+1} - T_s}{n+1} - (1 - p_2) \end{pmatrix} \xrightarrow{D} N_3(0, Diag(p_1, p_2 - p_1, 1 - p_2)),$$

where $T_k = \sum_{i=1}^{k} Y_i$.

Proof: Consider the representation, $\sqrt{n}(\frac{T_r}{n+1} - p_1) = \sqrt{n}(\frac{T_r}{n+1} - \frac{r}{n+1}) + \sqrt{n}(\frac{r}{n+1} - p_1)$. By CLT, the first term in the RHS of above converges in distribution to $N(0, p_1)$ whereas the second quantity converges ordinarily to 0 due to the assumption that $\sqrt{n}(r/n - p_1) \to 0$. Hence by Slutsky theorem $\sqrt{n}(\frac{T_r}{n+1} - p_1) \xrightarrow{D} N(0, p_1)$.

In a similar way the other convergences can also be proved using the CLT for iid random variables in combination with the facts that $\sqrt{n}(r/n - p_1) \to 0$ and $\sqrt{n}(s/n - p_2) \to 0$. However, the variables are independent exactly because they are based on disjoint sets of iid random variables. Hence the result follows.

Lemma 3: If $\sqrt{n}(r/n - p_1) \to 0$ and $\sqrt{n}(s/n - p_2) \to 0$ with $0 < p_1 < p_2 < 1$ then

$$\sqrt{n} \begin{pmatrix} U_{(r)} - p_1 \\ U_{(s)} - p_2 \end{pmatrix} \xrightarrow{D} N_2(0, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ & p_2(1-p_2) \end{pmatrix}).$$

Proof: First of all note that from Lemma 1,

$$\begin{pmatrix} U_{(r)} \\ U_{(s)} \end{pmatrix} \stackrel{D}{=} \begin{pmatrix} \frac{T_r}{T_{n+1}} \\ \frac{T_s}{T_{n+1}} \end{pmatrix} \stackrel{D}{=} \begin{pmatrix} \frac{T_r}{T_r + \overline{T_s - T_r} + \overline{T_{n+1} - T_s}} \\ \frac{T_r + \overline{T_s - T_r}}{T_r + \overline{T_s - T_r} + \overline{T_{n+1} - T_s}} \end{pmatrix}$$

Thus we define the vector valued function $\mathbf{g}(u_1, u_2, u_3) = (\frac{u_1}{u_1+u_2+u_3}, \frac{u_1+u_2}{u_1+u_2+u_3})$. Then it is easy to observe that

$$g(\frac{T_r}{n}, \frac{T_s - T_r}{n}, \frac{T_{n+1} - T_s}{n}) = (\frac{T_r}{T_{n+1}}, \frac{T_s}{T_{n+1}}) \stackrel{D}{=} \begin{pmatrix} U_{(r)} \\ U_{(s)} \end{pmatrix}.$$

Hence with already introduced notations, we get $G = \begin{pmatrix} 1-p_1 & -p_1 & -p_1 \\ 1-p_2 & 1-p_2 & -p_2 \end{pmatrix}$. Thus applying delta theorem we have $\sqrt{n}(U_{(r)}-p_1, U_{(s)}-p_2) \xrightarrow{D} N_2(0, GDiag(p_1, p_2-p_1, 1-p_2)G^T)$. A simple matrix multiplication gives $GDiag(p_1, p_2-p_1, 1-p_2)G^T = \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ & p_2(1-p_2) \end{pmatrix}$ and hence the result follows.

A generalization of Lemma 3: If $\sqrt{n}(r_i/n - p_i) \to 0, i = 1, 2, .., k$ with $0 < p_1 < p_2 < .. < p_k < 1$ then

$$\sqrt{n}(U_{(r_1)} - p_1, U_{(r_2)} - p_2, ..., U_{(r_k)} - p_k)^T \xrightarrow{D} N_k(0, \mathbf{p}(\mathbf{1} - \mathbf{p})^T),$$

where $\mathbf{p} = (p_1, p_2, .., p_k)^T$.

Now we are in a position, to derive our original result, that is, the asymptotic distribution of sample quantiles.

Asymptotic distribution of order statistics: If $X_{(i)}, i = 1, 2, .., n$, are order statistics from a continuous DF F having continuous density $F'(x) = f(x)$ such that $f(x)$ is positive in an nbd of population quantiles $\xi_{p_i}, i = 1, 2, .., k$ defined by $F(\xi_{p_i}) = p_i, i = 1, 2, .., k$ then

$$\sqrt{n}(X_{(r_1)} - \xi_{p_1}, X_{(r_2)} - \xi_{p_2}, ..., X_{(r_k)} - \xi_{p_k})^T \xrightarrow{D} N_k(0, \Sigma),$$

where $\Sigma = ((\sigma_{ij}))$ with $\sigma_{ij} = \frac{p_i(1-p_j)}{f(\xi_{p_i})f(\xi_{p_j})}$.

Proof: The proof of the above result follows from a simple application of delta method. Since by the assumption of the result, F is continuous and hence $F^{-1}(u)$ is well defined, we define the vector valued function $g(u_1, u_2, .., u_k) = (F^{-1}(u_1), F^{-1}(u_2), .., F^{-1}(u_k))^T$. Then

with the already introduced notations, G, the matrix with elements as differentials, is a diagonal matrix with elements $\frac{\partial F^{-1}(u_i)}{\partial u_i} = \{f(F^{-1}(u_i))\}^{-1}$. Hence

$$\sqrt{n}(X_{(r_1)} - \xi_{p_1}, X_{(r_2)} - \xi_{p_2}, ..., X_{(r_k)} - \xi_{p_k})^T \xrightarrow{D} N_k(0, G\mathbf{p}(\mathbf{1}-\mathbf{p})^T G^T).$$

A simple matrix multiplication yields the final result.

Asymptotic marginal distribution: With the already introduced notations and assumptions, for $r = [np+1]$, $\sqrt{n}(X_{(r)} - \xi_p)^T \xrightarrow{D} N(0, \sigma^2)$, where $\sigma^2 = p(1-p)/f^2(\xi_p)$.

2 Asymptotic distribution of quantile based summary measures

Now we shall derive the asymptotic joint distribution of quantile based summary measures like sample quartile deviation and median. or a sample of size n, sample quartile deviation is defined as $\hat{Q} = (X_{(s)} - X_{(r)})/2$, where $r = [n/4+1]$ and $s = [3n/4+1]$. If Q denotes the population QD, then it follows from the joint asymptotic distribution of sample order statistics that $\sqrt{n}(\hat{Q} - Q) \xrightarrow{D} N(0, \tau^2)$, where $\tau^2 = \frac{1}{64}(3/f^2(\xi_1) + 3/f^2(\xi_3) - 2/\overline{f(\xi_1)f(\xi_3)})$. However, for a symmetric distribution, $f(\xi_1) = f(\xi_3)$, and hence τ^2 reduces to $\{16f^2(\xi_1)\}^{-1}$. Again for a sample of size n, sample median is defined as $M_n = X_{(r)}$, where $r = [n/2+1]$ If M denotes the population QD, then it follows from the asymptotic distribution of sample order statistics that $\sqrt{n}(M_n - M) \xrightarrow{D} N(0, \tau^2)$, where $\tau^2 = \frac{1}{4f^2(\xi_2)}$.

3 Asymptotic distribution of extreme order statistics

We have assumed so far that $\sqrt{n}(r/n - p) \to 0$ for $p \in (0,1)$. But for extreme order statistics $(e.g. r = 1, n-1, n)$, $\frac{r}{n} \to 0$ or 1. Thus we need to develop the corresponding asymptotic distribution separately.

For a general development, first assume that $k/n \to 0$, that is k is fixed. It is already proved

in Lemma 1 that for fixed k, $U_{(k)} \stackrel{D}{=} \frac{T_k}{T_{n+1}}$, where $T_k \stackrel{D}{=} \sum_{i=1}^{k} Y_i$ for iid exponential(mean=1) random variables Y_i. Since, Y_i are iid exponential(mean=1) random variables, we have by WLLN, $\frac{T_{n+1}}{n} \stackrel{P}{\rightarrow} 1$. Hence from Slutsky Theorem $nU_{(k)} = \frac{T_k}{T_{n+1}/n} \stackrel{D}{\rightarrow} T_k$. Since $T_k \sim Gamma(rate = 1, shape = k)$, we have $nU_{(k)} \stackrel{D}{\rightarrow} Gamma(rate = 1, shape = k)$.

Next assume that $k/n \rightarrow 1$, that is $k = n, n-1$ for example.In particular consider $k = n$. Then from Lemma 1, $U_{(n)} \stackrel{D}{=} \frac{T_n}{T_{n+1}}$, where $T_k \stackrel{D}{=} \sum_{i=1}^{k} Y_i$ for iid exponential(mean=1) random variables Y_i. Hence $n(1 - U_{(n)}) \stackrel{D}{=} n(1 - \frac{T_n}{T_{n+1}}) = \frac{Y_{n+1}}{T_{n+1}/n}$. As earlier, it follows from WLLN that $\frac{T_{n+1}}{n} \stackrel{P}{\rightarrow} 1$. Hence from Slutsky Theorem $n(1 - U_{(n)}) \stackrel{D}{\rightarrow} Gamma(rate = 1, shape = 1)$. In a similar way, for $k = n - 1$, $n(1 - U_{(n-1)}) \stackrel{D}{=} n(1 - \frac{T_{n-1}}{T_{n+1}}) = \frac{Y_n + Y_{n+1}}{T_{n+1}/n} \stackrel{D}{\rightarrow} Gamma(rate = 1, shape = 2)$.

4 Asymptotic distribution of sample range & midrange for uniform parent

We start with the joint asymptotic distribution of extreme order statistics for uniform samples. Specifically, if $U_i, i = 1, 2, .., n$ are iid R(0,1) variables, we require the joint asymptotic distribution of $U_{(1)}$ and $U_{(n)}$. Take $a_n > 0$ and $b_n > 0$ and consider $P(a_n U_{(1)} > x \cap b_n(1 - U_{(n)}) > y)$.
Since $P(a_n U_{(1)} > x \cap b_n(1 - U_{(n)}) > y) = P(\frac{x}{a_n} < U_i < 1 - \frac{y}{b_n}, i = 1, 2, .., n)$, we get the final expression as $(1 - \frac{y}{b_n} - \frac{x}{a_n})^n$ provided $0 < \frac{x}{a_n} < 1 - \frac{y}{b_n} < 1$. Thus it seems sensible to take $a_n = b_n = n$ and as $n \rightarrow \infty$, the above inequality is satisfied for positive x and y.
Thus for positive x and y, $\lim_{n\rightarrow\infty}(1 - \frac{y}{n} - \frac{x}{n})^n = e^{-x-y}$. Hence $\begin{pmatrix} nU_{(1)} \\ n(1 - U_{(n)}) \end{pmatrix} \stackrel{D}{\rightarrow} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_i, i = 1, 2$ are iid exponential with mean unity.

Next we proceed to the asymptotic distribution of sample range and midrange. Let $U_i, i = 1, 2, .., n$ are iid R(0,1) variables and we want to obtain the asymptotic distribution of sample range $R_n = U_{(n)} - U_{(1)}$ and mid range $M_n = (U_{(n)} + U_{(1)})/2$. We have already

derived that for uniform sample, $\begin{pmatrix} nU_{(1)} \\ n(1-U_{(n)}) \end{pmatrix} \xrightarrow{D} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_i, i = 1, 2$ are iid exponential with mean unity. Hence using delta theorem with $g(x, y) = x + y$, we observe $nU_{(1)} + n(1 - U_{(n)}) = n(1 - R_n) \xrightarrow{D} Y_1 + Y_2 \sim Gamma(rate = 1, shape = 2)$. In a similar way $n(M_n - \frac{1}{2}) = \frac{1}{2}(nU_{(1)} - n(1 - U_{(n)}) \xrightarrow{D} \frac{1}{2}(Y_1 - Y_2) \sim DE(0, \frac{1}{2})$.

Large sample Inference: Module 18[1]

What we provide in this module

- Applying asymptotic results in inference
- Variance stabilizing transformations
- Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Large sample inference

We have discussed so far, asymptotic distributions of different summary measures. These results form the basis of developing different inferential procedures concerning different population characteristics. Starting from a general framework, we discuss both single sample and two sample procedures and assess their performance.

1.1 Estimation in large samples

Suppose, we have a large number of observations from a population indexed by the unknown parameter θ(may be vector valued). Then one may be naturally interested in providing a good(say, MVUE) estimate of θ. However, a good estimator is not easy to obtain, particularly when we don't have much idea about the parent population. However, if we confine ourselves to the class of estimators, which are very close to the actual parameter for sufficiently large n, we can use results of large sample procedures. For example, suppose we find a sequence of estimators $T_n, n \geq 1$ such that $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$ then $T_n \xrightarrow{P} \theta$ and hence T_n is a reasonable estimator of θ. Naturally, T_n may perform poorly in small samples.

1.2 Large sample tests & confidence interval estimation

Next suppose we are interested in $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ and only a large number of observations from a population indexed by the single unknown parameter θ, are available. Naturally, the usual methods like Neyman-Pearson lemma or Likelihood ratio methods can't be adopted as the underlying population is not specified. Suppose we find a sequence of statistics $T_n, n \geq 1$ satisfying $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$. Then a large sample procedure rejects the null hypothesis at an approximate level α if $\frac{\sqrt{n}(T_n - \theta_0)}{\sigma(\theta_0)} > \tau_{\alpha/2}$. An associated confidence interval with approximate confidence coefficient $1 - \alpha$ is easy to obtain from the set $\{\theta : |\frac{\sqrt{n}(T_n - \theta)}{\sigma(\theta)}| \leq \tau_{\alpha/2}\}$.

1.3 Problems regarding mean for single univariate population

Suppose n observations are available from a population with mean μ and finite variance σ^2. If we are interested in estimating μ then a consistent estimator is simply $\bar{X}$. Suppose we are interested in $H_0 : \mu = \mu_0$. For known σ, one can use $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ and for unknown σ, we use $\sqrt{n}(\bar{X} - \mu_0)/s$. For large n, each variable has a N(0,1) distribution under the null hypothesis. Hence a large sample test can be constructed. Next assume σ unknown, then to get a confidence interval for μ we start from the relation $P_\mu\{\sqrt{n}(\bar{X} - \mu)/s\} \approx 1 - \alpha$ and arrive at the interval $[\bar{X} - \tau_{\alpha/2}s/\sqrt{n}, \bar{X} + \tau_{\alpha/2}s/\sqrt{n}]$ with approximate coverage probability $1 - \alpha$.

1.4 Problems regarding variance for single univariate population

Suppose we are interested in estimating σ^2 or σ. Then consistent estimators are simply sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and s. Suppose we are interested in $H_0 : \sigma = \sigma_0$. Assume μ is unknown and parent distribution as normal. Then we use either $\sqrt{2n}(s - \sigma_0)/\sigma_0$ or $\sqrt{n/2}(s^2 - \sigma_0^2)/\sigma_0^2$, which are asymptotically N(0,1) under the null hypothesis. Hence large sample tests can be constructed. However, if the parent is not normal, then $\sqrt{n}(m_2 - \mu_2) \xrightarrow{D} N(0, \tau^2 = \mu_4 - \mu_2^2)$. Since m_4 is a consistent estimator of μ_4, we can use the statistic $\sqrt{n}\frac{(m_2 - \sigma_0^2)}{\hat{\tau}}$, where $\hat{\tau}$ is a consistent estimator of τ. Naturally, such a statistic is asymptotically N(0,1), and hence tests can be constructed as earlier. Using either statistic, one can construct the confidence interval with approximate coverage probability $1 - \alpha$.

1.5 Problems regarding mean for two univariate populations

Suppose n_i observations are available from a population with mean μ_i and finite variance $\sigma_i^2, i = 1, 2$. Assume that the two populations are independent. For $\mu = \mu_1 - \mu_2$, a consistent estimator is simply $\bar{X}_1 - \bar{X}_2$. Next suppose we are interested in $H_0 : \mu_1 - \mu_2 = \mu_0$. If $\sigma_i's$ are known, one can use $T_n = \frac{(\bar{X}_1 - \bar{X}_2 - \mu_0)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$. If, we assume that $n_1/(n_1 + n_2) \to \lambda \in (0, 1)$ and $n_2/(n_1 + n_2) \to 1 - \lambda$, then it follows that $\sqrt{n_1 + n_2}(\bar{X}_1 - \mu_1, \bar{X}_2 - \mu_2) \xrightarrow{D} N_2(0, Diag(\sigma_1^2/\lambda +$

$\sigma_2^2/\overline{1-\lambda}))$. Thus under the null hypothesis, T_n is asymptotically N(0,1) and hence large sample tests can be constructed.

1.6 Problems regarding mean for two univariate populations

However, if $\sigma_i's$ are unknown, we replace them by their consistent estimates $s_i^2, i = 1, 2$. Then applying Slutsky's theorem $T_n' = \frac{(\bar{X}_1 - \bar{X}_2 - \mu_0)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ is asymptotically normal and hence tests can be constructed. In addition, considering $\frac{(\bar{X}_1 - \bar{X}_2 - \mu)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ as a pivot for μ, one can find an approximate confidence interval $[(\bar{X}_1 - \bar{X}_2) - \frac{\tau_{\alpha/2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, (\bar{X}_1 - \bar{X}_2) + \frac{\tau_{\alpha/2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}]$ with approximate coverage probability $1 - \alpha$.

1.7 Problems regarding variance for two univariate populations

Suppose n_i observations are available from a population with unknown mean μ_i and finite variance $\sigma_i^2, i = 1, 2$. Assume that the two populations are independent. For σ_1/σ_2, a consistent estimator is simply s_1/s_2. Next suppose we are interested in $H_0 : \sigma_1^2 - \sigma_2^2 = 0$. For simplicity, we assume normal parents. Then considering the fact that $\sqrt{n_i}(s_i^2 - \sigma_i^2) \xrightarrow{D} N(0, 2\sigma_i^2)$ independently for each i, we consider $T_n = \frac{(s_1^2 - s_2^2 - 0)}{\sqrt{2\sigma^4/n_1 + 2\sigma^4/n_2}}$, where σ^2 is the common unspecified value under the null hypothesis. Thus we replace σ^2 by its pooled estimator $s^2 = (n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2)$. Since, under $n_1/(n_1+n_2) \to \lambda \in (0, 1)$ and $n_2/(n_1+n_2) \to 1-\lambda$, s^2 is consistent for σ^2, it follows that $T_n = \frac{(s_1^2 - s_2^2)}{\sqrt{2s^4/n_1 + 2s^4/n_2}} \xrightarrow{D} N(0, 1)$ under the null hypothesis and hence large sample tests can be constructed.

Again considering the fact that $\sqrt{n_i}(s_i - \sigma_i) \xrightarrow{D} N(0, \sigma_i^2/2)$ independently for each i, we can alternatively consider $T_n' = \frac{(s_1 - s_2 - 0)}{\sqrt{\sigma^4/\overline{2n_1} + \sigma^4/\overline{2n_2}}}$, where σ^2 is the common unspecified value under the null hypothesis. As earlier, we can replace σ^2 by its pooled estimator s^2. Then as before, $\frac{(s_1 - s_2)}{\sqrt{s^4/\overline{2n_1} + s^4/\overline{2n_2}}} \xrightarrow{D} N(0, 1)$ under the null hypothesis and hence large sample tests can be constructed. However, if normality is not assumed, then we should use the fact that $T_n' = \frac{(s_1^2 - s_2^2)}{\hat{SE}(s_1^2 - s_2^2)} \xrightarrow{D} N(0, 1)$, where moment based estimators are used to estimate the standard error. Further considering $\frac{(s_1^2 - s_2^2 - \overline{\sigma_1^2 - \sigma_2^2})}{\sqrt{2s^4/n_1 + 2s^4/n_2}}$ as a pivot for $\sigma_1^2 - \sigma_2^2$, one can find an approximate

confidence interval. In a similar way considering $\frac{(s_1-s_2-\overline{\sigma_1-\sigma_2})}{\sqrt{s^4/2n_1+s^4/2n_2}}$ as a pivot for $\sigma_1-\sigma_2$, the associated confidence interval can be obtained.

1.8 Problems regarding coefficient of variation

Suppose n iid observations from a normal population with unknown parameters are available. Denote the population cv and sample cv by V and v, respectively. Naturally a consistent estimator of V is v. Suppose, we want to perform a test for $H_0 : V = V_0$. Then looking at the fact that Again considering the fact that $T_n = \sqrt{n}(v-V) \xrightarrow{D} N(0, \frac{V^2}{2}(1+10^{-4}2V^2))$, we use the statistic $T_n = \sqrt{n}\frac{(v-V_0)}{\sqrt{\frac{V_0^2}{2}(1+10^{-4}2V_0^2)}}$, which under the null hypothesis is distributed as N(0,1). Thus we reject the null hypothesis if $T_n > \tau_\alpha$ or $T_n < -\tau_\alpha$ or $|T_n| > \tau_{\alpha/2}$ according as the alternative is $H : V > V_0$ or $H : V < V_0$ or $H : V \neq V_0$.For an approximate confidence interval for V, we start from the relation $P\{|\sqrt{n}\frac{(v-V)}{\frac{V^2}{2}(1+10^{-4}2V^2)}| \leq \tau_{\alpha/2}\} \approx 1-\alpha$ to get the confidence interval $[v - \tau_{\alpha/2}\sqrt{\frac{v^2}{2}(1+10^{-4}2v^2)}, v + \tau_{\alpha/2}\sqrt{\frac{v^2}{2}(1+10^{-4}2v^2)}]$.

1.9 Large sample tests for normality

Suppose n iid observations from a population with finite moments up to order eight. Often we are interested in testing the null hypothesis that the underlying distribution is normal. Suppose γ_1 and γ_2 are, respectively, the coefficient of skewness and kurtosis. Then it is well known that for normal parent $\gamma_1 = 0$ and $\gamma_2 = 0$. Thus, we can perform tests of normality via $H_{01} : \gamma_1 = 0$ and $H_{02} : \gamma_2 = 0$. For testing H_{01}, we use the statistic $\sqrt{n/6}g_1$, which is asymptotically N(0,1) under the null hypothesis. However, for testing H_{02}, we use the statistic $\sqrt{n/24}g_2$, which is asymptotically N(0,1) under the null hypothesis. For making the first test n=30 is sufficient to reach normality whereas for the second test n=200 ensures normality. But, these are tests for skewed and symmetric alternatives, respectively, and hence one can consider the combined hypothesis $H_{03} : \gamma_1 = 0\ , \gamma_2 = 0$. Then an obvious statistic is $n(g_1^2/6) + n(g_2^2/24)$ which under the null hypothesis and normality, is distributed asymptotically as χ_2^2.

2 Variance stabilization

2.1 Why stabilize ?

Suppose iid observations from k populations indexed by parameters $\theta_i, i = 1, 2, .., k$. Suppose a consistent estimator $T_{in}, i = 1, 2, .., k$ is available. Then often the asymptotic variance $\sigma^2(\theta_i)$ depends on θ_i and we need to replace them by consistent estimates for making a test. Such substitution makes the convergence slower. Thus, it would be better if we can use some transformation to make the asymptotic variance independent of the parameter.

Thus we need some transformation $h(T)$ for some continuous function h with a non zero derivative at θ, such that the asymptotic variance $\{h'(\theta)\}^2\sigma^2(\theta)$ is independent of θ. Hence we need to decide h such that $\{h'(\theta)\}^2\sigma^2(\theta) = c^2$ is satisfied for some c independent of θ. Thus h must satisfy the differential equation $\frac{dh(\theta)}{d\theta} = \frac{c}{\sigma(\theta)}$ so that h can be determined up to a change of origin and scale.

2.2 Examples

Binomial proportion: Suppose $X_i, i = 1, 2, ..$ are iid Bin(1,p) variables. Then $\sqrt{n}(\bar{X} - p) \xrightarrow{D} N(0, p(1-p))$. We are looking for some transformation $h(\bar{X})$ making the asymptotic variance constant. Then $h(p) = c\int \frac{dp}{\sqrt{p(1-p)}} + k$, which on integration gives $h(p) = 2csin^{-1}(\sqrt{p}) + k$. Choosing $c = .5$ and $k = 0$, we get the transformation $h(p) = sin^{-1}(\sqrt{p})$.

Poisson mean: Suppose $X_i, i = 1, 2, ..$ are iid $Poisson(\theta)$ variables. Then $\sqrt{n}(\bar{X} - \theta) \xrightarrow{D} N(0, \theta^2)$. We are looking for some transformation $h(\bar{X})$ making the asymptotic variance constant. Then $h(p) = c\int \frac{d\theta}{\sqrt{\theta}} + k$, which on integration gives $h(\theta) = 2c\sqrt{\theta} + k$. Choosing $c = .5$ and $k = 0$, we get the transformation $h(\theta) = \sqrt{\theta}$.

Normal variance: Suppose $X_i, i = 1, 2, ..$ are iid $N(\mu, \sigma^2)$. Then $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4)$. As earlier, we need some transformation $h(s^2)$ making the asymptotic variance independent of σ^2. Then $h(\sigma^2) = c\int \frac{d\sigma^2}{\sqrt{2\sigma^4}} + k$, which on integration gives $h(\sigma^2) = \frac{c}{\sqrt{2}} log\sigma^2 + k$.

Choosing $c = \sqrt{2}$ and $k = 0$, we get $h(\sigma^2) = log\sigma^2$.

Correlation coefficient: Suppose $(X_i, Y_i), i = 1, 2, ..$ be iid observations from a bivariate distribution with correlation coefficient ρ. Then for the sample correlation coefficient r, it is already deduced that $\sqrt{n}(r - \rho) \xrightarrow{D} N(0, (1 - \rho^2)^2)$. As earlier, we need some function h making the asymptotic variance independent of the parameter. Then as earlier, $h(\rho) = c \int \frac{d\rho}{1-\rho^2} + k$, which on integration gives $h(\rho) = \frac{c}{2} log(\frac{1+\rho}{1-\rho}) + k$. Choosing $c = 1$ and $k = 0$, we get the desired transformation $h(\rho) = \frac{1}{2} log(\frac{1+\rho}{1-\rho}) = tanh^{-1}\rho$. This particular transformation is known as Fisher's Z transformation.

2.3 Convergence rate of the variance stabilized transformation

The greatest advantage of variance stabilizing transformations is that normality is reached more rapidly than the original statistic. As an example, consider Fisher's z transformation. It is already noted that $\sqrt{n}(r - \rho) \xrightarrow{D} N(0, (1 - \rho^2)^2)$. Then from Slutsky theorem, it follows that $T_{1n} = \frac{\sqrt{n}(r-\rho)}{1-r^2} \xrightarrow{D} N(0, 1)$. Again for Fisher's Z transformation $T_{2n} = \sqrt{n}(tanh^{-1}r - tanh^{-1}\rho) \xrightarrow{D} N(0, 1)$. Thus for comparison, we conduct a simulation study with bivariate normal distribution having correlation coefficient ρ and provide a histogram together with the normal density plot for both the variables for different choices of n and ρ. All these can be found in the next pages.

It is observed that for both the variables the convergence to normality is reached for ρ near 0. The transformed variable converges at a faster rate. However, for very large and very small values of ρ, the convergence to normality is not seen even for n=150 for the original and for n=100 for the transformed. In particular for ρ nearby ± 1, the distribution is quite different from the standard normal for large n.

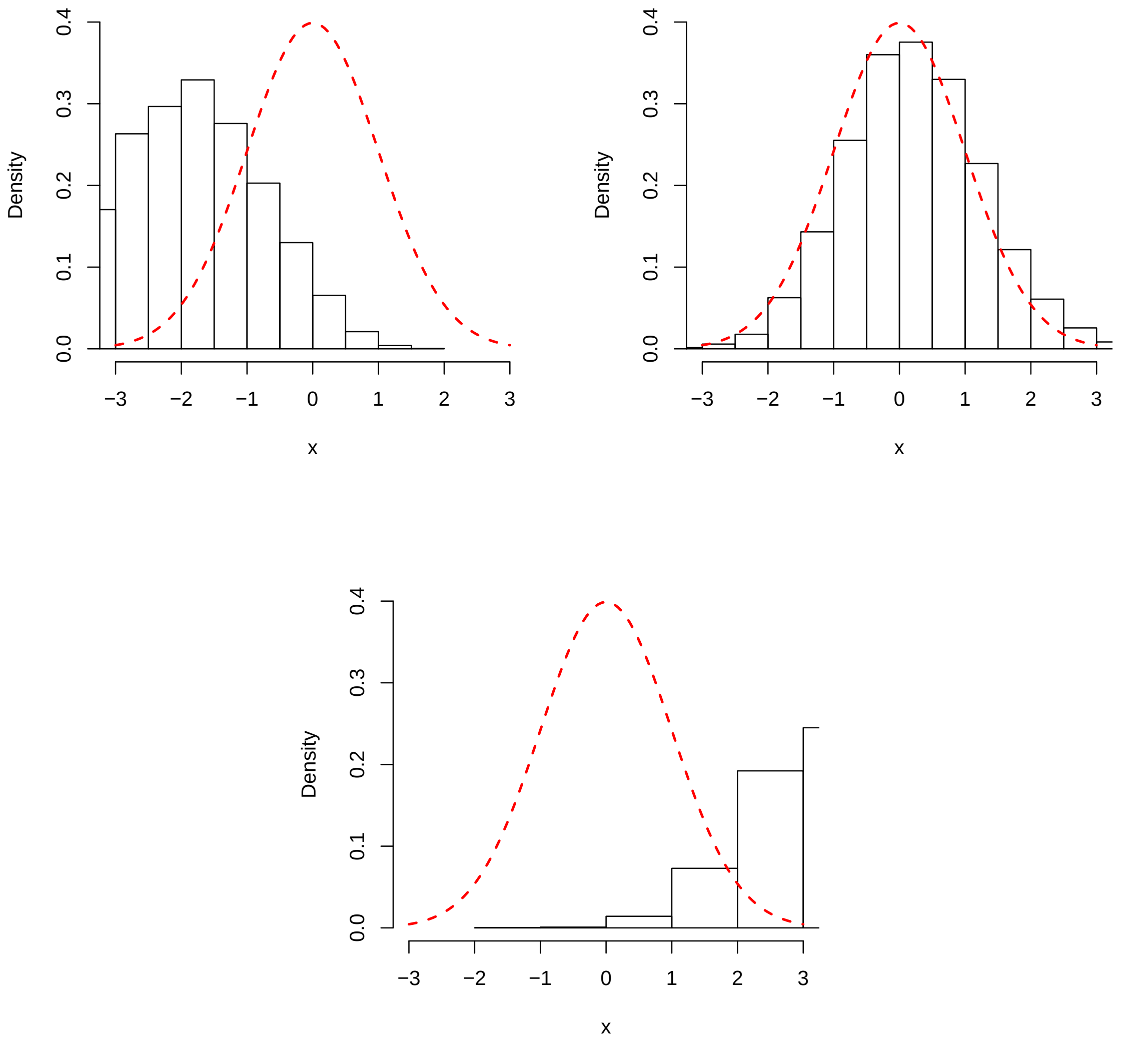


Figure 1: **Convergence rate of T_1 for $\rho = -.6, .3, .7$ with n=150(clockwise)**

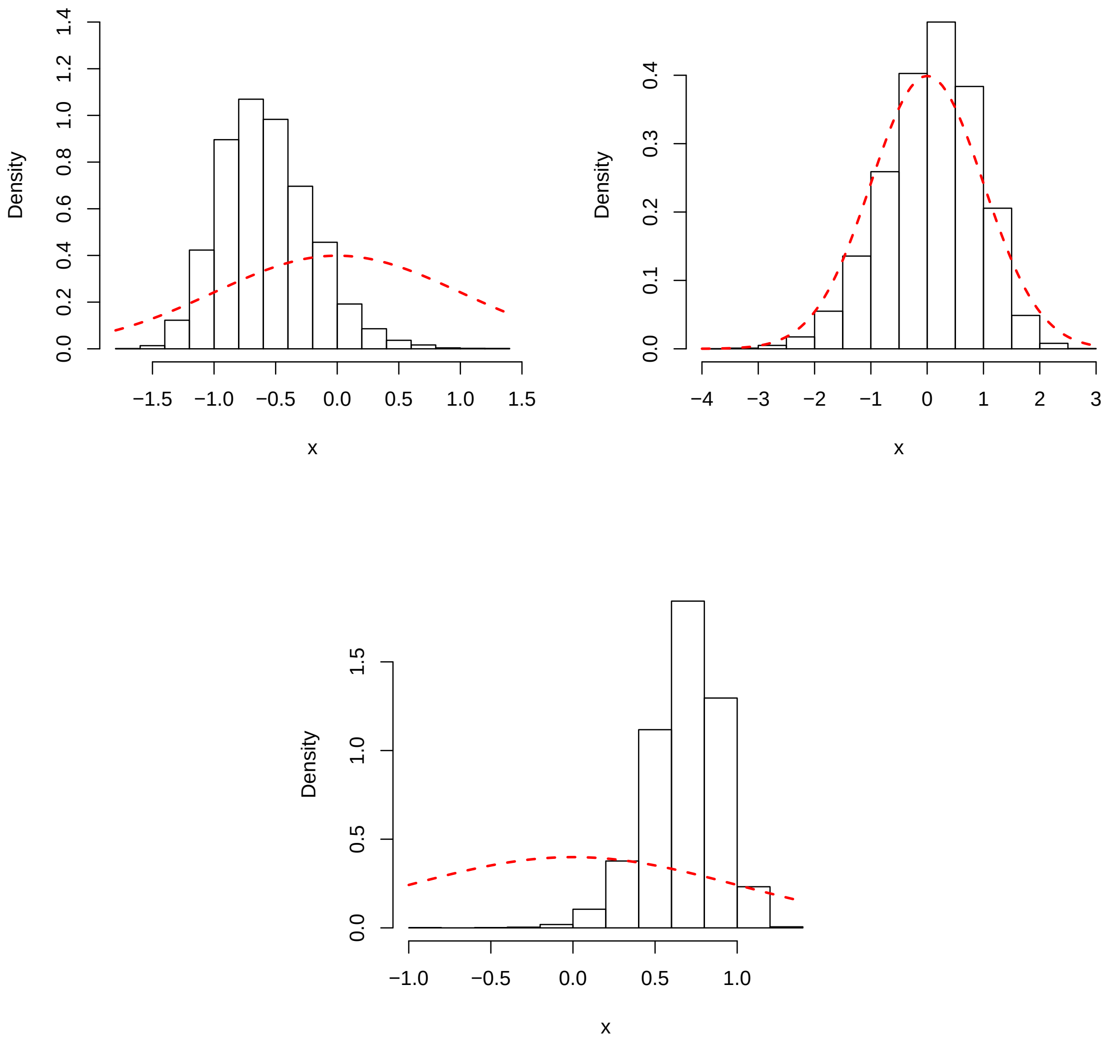


Figure 2: **Convergence rate of T_2 for $\rho = -.6, .3, .7$ with n=100(left to right)**

Large sample Inference: Module 19[1]

What we provide in this module

- Applications of variance stabilization for k populations
- Pearson chi-square tests
- Applications

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Use of Z transformation for bivariate populations

1.1 Inferential procedures for a single population

Suppose $(X_i, Y_i), i = 1, 2, ..$ are iid observations from a bivariate normal distribution with correlation coefficient ρ. Suppose, we want to test $H_0 : \rho = \rho_0$ for $\rho \neq 0$. Then we can use the statistic $T_n = \sqrt{n}(tanh^{-1}r - tanh^{-1}\rho_0)$, which under the null hypothesis is asymptotically $N(0,1)$. Thus we reject H_0 against $H_1 : \rho \neq \rho_0$ if $|T_n| > \tau_{\alpha/2}$. Similarly the one sided hypothesis can also be considered. Again considering $\sqrt{n}(tanh^{-1}r - tanh^{-1}\rho_0)$ as a pivot, one can obtain a confidence interval $[tanh(z_1), tanh(z_2)]$ with approximate confidence coefficient $1-\alpha$, where $z_1 = tanh^{-1}(r) - \tau_{\alpha/2}/\sqrt{n}$ and $z_2 = tanh^{-1}(r) + \tau_{\alpha/2}/\sqrt{n}$.

1.2 Use of Z transformation for k bivariate populations

Estimation: Suppose we k independent samples from k bivariate normal populations with common correlation coefficient ρ. Assume that n_i observations are available from the i th population and r_i is the sample correlation coefficient for the i th population. Suppose it is required to obtain a pooled estimate of ρ. Define $z_i = tanh^{-1}r_i$, then asymptotically it has a normal distribution with mean $\zeta = tanh^{-1}\rho$ and variance $1/n_i$. Hence considering the inverse variance as the weights, we get the pooled estimator of ζ as $\hat{\zeta} = \frac{\sum_{i=1}^k n_i z_i}{\sum_{i=1}^k n_i}$. Thus a pooled estimator of ρ can be obtained from the above as $\hat{\rho} = tanh(\hat{\zeta}) = \frac{exp(2\hat{\zeta})-1}{exp(2\hat{\zeta})+1}$. However, an improved estimator can be obtained if we replace n_i by $n_i - 3$.

Hypothesis testing: Suppose independent samples of size n_i are available from k bivariate normal populations with correlation coefficient ρ_i for the i th population. Suppose we are interested in $H_0 : \rho_1 = \rho_2 = .. = \rho_k$. If r_i is the sample correlation coefficient for the i th population, then $\sqrt{n_i}(z_i - \zeta_i)$ is asymptotically N(0,1). If ζ_0 is the common unspecified value under the null hypothesis, then we may take $T_n = \sum_{i=1}^k (n_i - 3)(z_i - \zeta_0)^2$ which is approximately χ^2_k. However ζ_0 is not specified and hence we replace it by pooled estimator $\hat{\zeta}_0 = \frac{\sum_{i=1}^k (n_i-3)z_i}{\sum_{i=1}^k (n_i-3)}$ to get the modified statistic $T_n' = \sum_{i=1}^k (n_i - 3)(z_i - \hat{\zeta}_0)^2$ which

is approximately χ^2_{k-1}. Thus a large sample test would reject the null hypothesis if the observed value of T_n' exceeds $\chi^2_{k-1,\alpha}$.

The same methods can be adopted for the homogeneity of k Bernoulli population, or Normal population with respect to variance or Poisson population. In, addition, for k independent samples of different sizes from the same population, one can , as earlier obtain pooled estimates of the parameter of interest. However, the subsequent development is routine and hence we skip the details.

2 Asymptotic distributions related to quadratic forms

Often our test statistic comes in the form of a quadratic form. If the exact distribution of such a statistic is not tractable, asymptotic methods are the only options. However, the asymptotic distribution of a quadratic form is not the usual normal but often it is chisquare. Thus we start with some basic asymptotic results on quadratic forms and later provide applications in statistical hypothesis testing.

2.1 Few results

Lemma 1: If $\mathbf{X}^{pX1} \sim N_p(\mu, \Sigma)$, with positive definite Σ, then $Z = (\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu) \sim \chi^2_p$.

Proof: Since Σ is positive definite, there exists a non singular matrix C such that $C\Sigma C^T = I_p$. Now, if we define $\mathbf{Y} = C(\mathbf{X}\text{-}\mu)$, then $\mathbf{Y} \sim N_p(0, I_p)$. Now $Z = \mathbf{Y}^T\mathbf{Y} = \sum_{i=1}^p Y_i^2$, where Y_i are iid standard normal variates. Hence by definition, $Z \sim \chi^2_p$.

Lemma 2: If $\mathbf{X}_i, i \geq 1$ are iid p component random vectors with mean μ and dispersion matrix $\Sigma > 0$. Then

$$W = n(\bar{X} - \mu)^T S_n^{-1} (\bar{X} - \mu) \xrightarrow{D} \chi^2_p,$$

where $S_n = n^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{X})(\mathbf{X}_i - \bar{X})^T$ is the sample variance covariance matrix.

Proof: Using multivariate CLT, we have $\sqrt{n}(\bar{X}-\mu) \xrightarrow{D} \mathbf{Y}$, where $\mathbf{Y} \sim N_p(0,\Sigma)$. Since, $S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{X})(\mathbf{X}_i - \bar{X})^T = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T - \bar{X}(\bar{X})^T$, it follows from WLLN that $S_n \xrightarrow{P} (\Sigma + \mu\mu^T) - \mu\mu^T = \Sigma$. Hence from Slutsky theorem, $W \xrightarrow{P} \mathbf{Y}^T\Sigma^{-1}\mathbf{Y}$. Thus applying Lemma 1 above, we get $\mathbf{Y}^T\Sigma^{-1}\mathbf{Y} \sim \chi_p^2$ and hence the result follows.

3 Pearsonian χ^2 statistic

3.1 The statistic & the asymptotic distribution

Suppose n independent trials are performed. Each trial has k mutually exclusive and exhaustive outcomes $A_1, A_2, .., A_k$ with p_j as the probability for the j th outcome. Naturally $\sum p_j = 1$. Let f_j be the number of trials with outcome A_j so that $\sum f_j = n$. Then Pearson's χ^2 statistic is defined as

$$T_n = \sum_{i=1}^{k} \frac{(f_j - np_j)^2}{np_j}.$$

Now we deduce the asymptotic distribution of such a statistic.

Result: As $n \to \infty$,

$$T_n \xrightarrow{D} \chi_{k-1}^2.$$

Proof: Define the k component outcome vector for the i th trial $X^{(i)}$, where $P(X_j^{(i)} = 1) = p_j$ and $P(X_j^{(i)} = 0) = 1 - p_j$ so that exactly k-1 components of $X^{(i)}$ are zeros and a one such that $\sum_{j=1}^{k} X_j^{(i)} = 1$. Then $X^{(i)}, i = 1, 2..$ are iid vectors with $X^{(i)} \sim Multinomial(n, p_1, .., p_k), i = 1, 2, .., n$ so that $E(X^{(i)}) = \mathbf{p} = (p_1, p_2, .., p_k)^T$ and $Disp(X^{(i)}) = Diag(p_1, p_2, .., p_k) - \mathbf{p}\mathbf{p}^T$. However $Disp(X^{(i)})$ is singular and hence, we introduce $Y^{(i)} = (X_1^{(i)}, X_2^{(i)}, .., X_{k-1}^{(i)})^T, i = 1, 2, .., n$, then $Y^{(i)}$ are iid vectors with $EY^{(i)} = (p_1, p_2, .., p_{k-1})^T$ and $Disp(Y^{(i)}) = \Sigma = Diag(p_1, p_2, .., p_{k-1}) - (p_1, p_2, .., p_{k-1})^T(p_1, p_2, .., p_{k-1})$.
Since Σ is positive definite, it follows from multivariate CLT that $Z_n = \sqrt{n}(\bar{\mathbf{Y}} - (p_1, p_2, .., p_{k-1})) \xrightarrow{D} N_{k-1}(0, \Sigma)$. Hence, it follows from Lemma 2 that $Z_n^T\Sigma^{-1}Z_n \xrightarrow{D} \chi_{k-1}^2$. Now $\Sigma^{-1} = Diag(1/p_1, 1/p_2, ..,$ $\frac{1}{p_k}\mathbf{1}\mathbf{1}^T$, where $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Then noting the fact that $\bar{\mathbf{Y}} = (f_1/n, f_1/n, .., f_{k-1}/n)^T$ and

$f_k = n - \sum_{i=1}^{k-1} f_i$, a simple algebra with quadratic form expresses $Z_n^T \Sigma^{-1} Z_n$ as T_n. This completes the result.

3.2 Applications of Pearson's χ^2 statistic

3.2.1 Tests for goodness of fit

Often the underlying distribution of the data is not known and traditional practice is to use normality. But such an assumption makes the inference weaker. Therefore the experimenter wish to know whether the underlying distribution is of a given form. That is, the objective is to know which distribution "fits" good to the observed data. Tests for determining the underlying distribution, which is a good fit to the observed data are called goodness-of-fit tests.

We start with a motivating example. Suppose the number of misprints per page of a book of 300 pages are represented in the form of a grouped distribution:

No. of misprints/page	0	1	2	3	≥ 4
Number of pages	200	84	13	3	0

Here the number of observations is quite large to assume normality. But the variable is discrete with small number of categories. It is well known that the number of misprints/page has, in general, a Poisson distribution. But the mean, i.e. the only parameter, is not specified in advance. Then how to test the hypothesis that the observations are coming from a Poisson distribution with unknown parameter.

For better understanding, consider another example. Consider the heights of 100 students in a certain class in the form of a grouped distribution:

Height(cm.)	154-159	159-164	164-169	169-174	174-179
# students	4	15	54	22	5

Here the number of observations is quite large to assume normality. However, the data is presented in the form of a frequency distribution. Moreover, the distribution of height is known to be normal. But he even under the assumption of normality, the parameters are not given. Then how to check the observations are coming from a normal distribution.
From these examples, we see that we often have data in the form of frequency distributions either from a known discrete distribution but with unspecified parameter(s) or from a known continuous distribution with unspecified parameter(s). In any situation the objective was to test whether a given distribution fits the data well.

Now we shall discuss such a procedure considering Pearson statistic. Consider n observations on a random variable X. Let the range of X be divided into mutually exclusive and exhaustive sets $A_i, i = 1, 2, .., k$ and observations are classified into these classes. Let f_i be the number of observations classified in A_i. Then $\sum_{i=1}^{k} f_i = n$. Define p_i as the probability that an X observation falls in A_i under the null hypothesis. Suppose we have a hypothesis $H_0 : p_1 = p_1^0, .., p_k = p_k^0$, where $p_i^0, i = 1, 2, ...k$ are all specified. Then the joint distribution of $(f_1, f_2, .., f_{k-1})$ is k-1 variate multinomial with parameters $(n, p_1, p_2, .., p_{k-1})$. Thus under the null hypothesis f_i are expected to be closer to np_i for each i. Then departure from the null hypothesis can be measured by Pearson's Chi-square statistic $T_n = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$. Now the higher is the discrepancy , the larger is the value of T_n and hence a higher value indicates departure from the null hypothesis. Since, under the null hypothesis $T_n' = \sum_{i=1}^{k} \frac{(f_i - np_i^0)^2}{np_i^0} \xrightarrow{D} \chi^2_{k-1}$, a large sample test rejects the null hypothesis if $T_n' > \chi^2_{k-1,\alpha}$.

Large sample Inference: Module 20[1]

What we provide in this module

- Applications of Pearson χ^2
- Consistency
- Comparing consistent estimators

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

1 Applications of Pearson's χ^2 statistic

1.1 Tests for goodness of fit

We have already discussed that a goodness of fit hypothesis takes the form $H_0 : p_1 = p_1^0, .., p_k = p_k^0$, where p_i^0 may be completely specified or unknown. For known p_i^0, we have seen how $U_k = \sum_{i=1}^k \frac{(f_i - np_i^0)^2}{np_i^0}$ can be used to measure the discrepancy. Since, under the null hypothesis $U_k = \sum_{i=1}^k \frac{(f_i - np_i^0)^2}{np_i^0} \xrightarrow{D} \chi^2_{k-1}$, a large sample test, which rejects the null hypothesis if $U_k > \chi^2_{k-1,\alpha}$ was suggested.

Now we shall discuss the corresponding procedure when $p_i's$ depend on unknown parameters. Suppose under the null hypothesis p_i depends on the unknown parameters $(\theta_1, \theta_2, .., \theta_r)$. Thus, in such a case $p_i = p_i(\theta_1, \theta_2, .., \theta_r)$, a function of the unknown parameters. Thus we need to estimate $\theta's$ and we suggest to use method of moments(MM). Suppose $\hat{\theta}_i, i = 1, 2, ..r$ be the MM estimates based on the data. If $\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, .., \hat{\theta}_r), i = 1, 2, .., k$ are the estimates of $p_i, i = 1, 2, .., k$, then Pearson's Chi-square statistic takes the form $V_k = \sum_{i=1}^k \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}$.

It can be shown, as earlier that under the null hypothesis, $V_k \xrightarrow{D} \chi^2_{k-r-1}$ as $n \to \infty$. Since a higher value of V_k indicates departure from the null hypothesis, a right tailed test based on V_k is appropriate. Thus, the large sample test rejects the null hypothesis if $V_k > \chi^2_{k-r-1,\alpha}$ at $100(1-\alpha)\%$ level of significance.

Often the data is available in the form of a grouped distribution. If np_i(or its estimate) is less than 5 for some i, the corresponding class is pooled with one or more neighbour classes so that the expected frequency for the combined class is at least 5. However, in such a case, degrees of freedom becomes number of classes after combining minus 1 minus the number of parameters estimated.

1.2 Tests for association

Suppose we have two characters A and B, where A has k classes $A_i, i = 1, 2, .., k$ and B has l classes $B_j, j = 1, 2, .., l$ and a population is classified according to these characters. Let

p_{ij} be the population proportion for the cell (A_i, B_j) then $\sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij} = 1$. Naturally, $p_{i0} = \sum_{j=1}^{l} p_{ij}$ is the population proportion for A_i and $p_{0j} = \sum_{i=1}^{k} p_{ij}$ is the population proportion for B_j. Suppose we are interested in $H_0 :$ A and B are independent or equivalently $H_0 : p_{ij} = p_{i0}p_{0j} \ \forall \ (i, j)$. Pearson's χ^2 statistic can be used for testing the above hypothesis of association.

First of all assume that (p_{i0}, p_{0j}) are all specified. Suppose n observations are taken from the above population and we denote the frequency of the cell (A_i, B_j) by f_{ij}. Then the joint distribution of $f_{ij}, i = 1, 2, .., k, j = 1, 2, .., l$ is multinomial with parameters n and $p_{ij}, i = 1, 2, .., k, j = 1, 2, .., l$. Then as before one can use the statistic $\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{ij}-np_{ij})^2}{np_{ij}}$ to measure the departure from the null hypothesis. Naturally, the higher is the discrepancy , the larger is the value of the statistic. Since, under the null hypothesis $T_n = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{ij}-np_{i0}p_{0j})^2}{np_{i0}p_{0j}} \xrightarrow{D} \chi^2_{kl-1}$, a large sample test rejects the null hypothesis if $T_n > \chi^2_{k-1,\alpha}$ provided (p_{i0}, p_{0j}) are all specified.

Most often the null hypothesis does not specify (p_{i0}, p_{0j}) and hence requires estimation from the data.For the observed data maximum likelihood estimators are $\hat{p}_{i0} = f_{i0}/n$ and $\hat{p}_{0j} = f_{0j}/n$. Plugging in these estimates, we get the statistic, $T'_n = n\{\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{f_{ij}^2}{f_{i0}f_{0j}} - 1\}$. Since, under the null hypothesis $T'_n \xrightarrow{D} \chi^2_{\overline{k-1}\ \overline{l-1}}$, a large sample test rejects the null hypothesis if $T'_n > \chi^2_{\overline{k-1}\ \overline{l-1},\alpha}$.

1.3 Tests for homogeneity

Suppose we have k populations classified in the same manner according to a character A, where A has l classes $A_j, j = 1, 2, .., l$. Let p_{ij} be the population proportion for the class A_j in the i th population. Thus $\sum_{j=1}^{l} p_{ij} = 1 \ \forall \ i$. We are interested in testing whether the k populations are similar(i.e. homogeneous). That is we want to test the null hypothesis $H_0 : p_{1j} = p_{2j} = .. = p_{kj} = p_j \ \forall \ j$ and Pearson's χ^2 statistic is also useful in this case.

Assume that p_j's are all specified. Suppose samples of size n_i is taken from the i th population,i=1,2,..,k. Define f_{ij} is the observed frequency for class A_j for the i th population so that

$\sum_{j=1}^{l} f_{ij} = n_i$. Then the joint distribution of $f_{ij}, j = 1, 2, .., l$ is multinomial with parameters n_i and $p_{ij}, j = 1, 2, .., l$ independently for each i. Thus $\sum_{j=1}^{l} \frac{(f_{ij}-n_i p_{ij})^2}{n_i p_{ij}}$ is approximately a χ^2_{l-1} variate. Due to independence to $\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{ij}-n_i p_{ij})^2}{n_i p_{ij}}$ is approximately a $\chi^2_{k\overline{l-1}}$. Since, under the null hypothesis $T_n = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{ij}-n_i p_j)^2}{n_i p_j} \xrightarrow{D} \chi^2_{k\overline{l-1}}$, and a higher value is an indicative of departure, a large sample test rejects the null hypothesis if $T_n > \chi^2_{k\overline{l-1},\alpha}$ provided p_j are all specified.

However, $p_j's$ are not always specified and we need to substitute them by appropriate estimates. One can use the polled estimate $\hat{p}_j = \frac{\text{Total frequency for the j th class}}{n} = \frac{\sum_{i=1}^{k} f_{ij}}{n}$ for the purpose. Plugging in these estimates, we get the statistic $T_n' = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{ij}-n_i\hat{p}_j)^2}{n_i\hat{p}_j} = n\{\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{f_{ij}^2}{n_i f_{0j}} - 1\}$, which under H_0 is asymptotically $\chi^2_{\overline{k-1}\ \overline{l-1}}$. Then, as earlier, a large sample test rejects the null hypothesis if $T_n' > \chi^2_{\overline{k-1}\ \overline{l-1},\alpha}$.

2 Yates correction for continuity

It is already mentioned that, we need expected frequencies or their estimates to be large enough for χ^2 approximation to hold. However, such a procedure can't be adopted for 2X2 tables. Yates suggested a modification, called Yates correction for continuity, for the Pearson χ^2 statistic to use in such 2X2 tables with low cell frequencies. The idea is that for $X \sim Bin(n,p)$ variable with large n, $\Phi(\frac{x+.5-np}{\sqrt{np(1-p)}})$ approximates $P(X \leq x)$ better than the usual approximation $\Phi(\frac{x-np}{\sqrt{np(1-p)}})$. In a similar way, $1-\Phi(\frac{x-.5-np}{\sqrt{np(1-p)}})$ approximates $P(X \geq x)$ better than the usual approximation $1-\Phi(\frac{x-np}{\sqrt{np(1-p)}})$. The idea is to make a smaller(larger) x slightly larger(smaller).
Consider a 2X2 contingency table with observed frequencies a,b,c and in the clockwise order with deficiency to qualify for the χ^2 approximation.

Yates correction suggests to add or subtract .5 to each cell frequency keeping marginal totals unchanged. For $ad > bc$, we consider the modified frequencies a-.5,b+.5,c+.5 and d-.5, whereas for $ad < bc$, the modified frequencies a+.5,b-.5,c-.5 and d+.5 are suggested.

Table 1: **2X2 Contingency Table**

	Counts	Counts	Total
Counts	a	b	a+b
Counts	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

Then the modified statistic χ^2 takes the forms $\frac{(ad-bc+n/2)^2}{(a+c)(b+d)((a+b)(c+d)}$ and $\frac{(ad-bc-n/2)^2}{(a+c)(b+d)((a+b)(c+d)}$, respectively.

3 Large sample properties of estimators

Unbiasedness, possessing minimum variance are small properties of estimators. But often the estimators takes implicit forms or even if the form is tractable exact sampling distribution is difficult to obtain. Thus we need alternative criterion for evaluating the performance of estimators. One such property is consistency. The idea is closely related to convergence concepts of probability.

3.1 Consistency

Let $X_i, i \geq 1$ be observations from a population indexed by a parameter θ. Suppose the interest is to obtain a good substitute of some real valued function $g(\theta)$. Let $T_n = T_n(X_1, .., X_n)$ be an estimator based on data. It is desirable to have $|T_n - g(\theta)|$ as close as possible to zero. But $|T_n - g(\theta)|$ is a random quantity and hence varies across samples. Thus, as an alternative, we need to keep the possibility of a low $|T_n - g(\theta)|$ as high as possible. Therefore, we need to maintain $P\{|T_n - g(\theta)| < \epsilon\}$ or $P\{\sup_{m\geq n} |T_m - g(\theta)| < \epsilon\}$ as high as possible. As n increases, more and more information is gathered and hence for a reasonable T_n one would expect $P\{|T_n - g(\theta)| < \epsilon\} \to 1$ or $P\{\sup_{m\geq n} |T_m - g(\theta)| < \epsilon\} \to 1$ for every $\epsilon > 0$. In the former case T_n is said to be weakly consistent and in the later case T_n is called strongly

consistent for θ. However, we confine ourselves to weakly consistent or simply consistent estimators.

3.1.1 Sufficient conditions for consistency

Let T_n be consistent for $g(\theta)$ and assume that T_n has a finite variance. Then from Chebyshev inequality, for any $\epsilon > 0$,

$$P(|T_n - g(\theta)| > \epsilon) \leq \epsilon^{-2} E\{T_n - g(\theta)\}^2$$

. Thus consistency of T_n holds if $MSE(T_n) = E\{T_n - g(\theta)\}^2 \to 0$. Now, we at once have the relation that

$$MSE(T_n) = Var(T_n) + \{E(T_n) - \theta\}^2.$$

Thus we have the equivalent conditions for consistency as $E(T_n) \to \theta$ and $Var(T_n) \to 0$.

3.1.2 The condition is not necessary

Suppose X_i are iid random variables with mean μ and finite variance σ^2. Define T_n such that $T_n = \bar{X}$ with probability $1 - 1/\sqrt{n}$ and $T_n = n$ with probability $1/\sqrt{n}$. First we shall show that $T_n \xrightarrow{P} \mu$. First of all note that for every $\epsilon > 0$,

$$P(|T_n - \mu| < \epsilon) \geq (1 - 1/\sqrt{n}) P(|\bar{X} - \mu| < \epsilon).$$

Since $\bar{X} \xrightarrow{P} \mu$, the first quantity in the RHS of above converges to unity. Again $1/\sqrt{n} \to 0$ and hence for every $\epsilon > 0$, $P(|T_n - \mu| < \epsilon) \to 1$. However, if F_n is the DF of $\bar{X}$, then we observe that the integral

$$\int |T_n| dF_n = (1 - 1/\sqrt{n}) \int |\bar{X}| dF_n + n/\sqrt{n}$$

diverges and hence the expected value and hence the MSE of T_n is not finite. Thus, we find that T_n is consistent but does not qualify for the sufficient condition.

3.2 Comparing consistent estimators: Concept of efficiency

3.2.1 MSE criterion

If T_{1n} and T_{2n} are both consistent for θ, then $MSE(T_{kn}) \to 0$ for k=1,2. Thus the estimator, which converges at a faster rate is preferable. For example, if we can show that $MSE(T_{1n}) < MSE(T_{2n})$, then T_{1n} converges at a faster rate.

Let us explain the criteria by means of an example. Consider estimation of σ^2 based on iid observations from a normal population with mean μ and variance σ^2. Suppose, we consider three estimators, $T_{kn} = c_{kn}s^2, k = 1, 2, 3$ with $c_{1n} = 1/n$, $c_{2n} = 1/\overline{n+1}$ and $c_{3n} = 1/\overline{n-1}$. Then T_{1n} is MLE, T_{2n} is minimum MSE estimator and T_{3n} is the MVUE. Considering the fact that $(n-1)s^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \bar(X))^2/\sigma^2 \sim \chi^2_{n-1}$, it is easy to establish the ordering $MSE(T_{2n}) < MSE(T_{1n}) < MSE(T_{3n})$. Hence T_{2n} converges at a faster rate. Thus it is interesting to note that the best estimator for small samples(i.e. MVUE) performs poorer in large samples.

3.2.2 Asymptotic variance criterion

However, finding exact MSE requires huge calculation and hence we develop a simpler criterion based on asymptotic normality. A consistent estimator T_n for $\gamma(\theta)$ is said to be consistent asymptotically normal(CAN), if in addition to consistency, asymptotic normality, that is, $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$ holds for some $\sigma^2(\theta) > 0$. Naturally among CAN estimators, the estimator with the smallest possible $\sigma^2(\theta) > 0$ is the best, that is, best asymptotically normal(BAN).

Looking at the fact that in a regular estimation case, based on n iid observations, the minimum attainable variance for an unbiased estimator of $\gamma(\theta)$ is $\frac{[\gamma'(\theta)]^2}{nI(\theta)}$, where $I(\theta)$ is the Fisher's information based on a single observation, Fisher argued that the minimum attainable value of $\sigma^2(\theta)$ is $\frac{[\gamma'(\theta)]^2}{I(\theta)}$. Consequently, Fisher defined T_n as asymptotically efficient/optimal if $\sigma^2(\theta) = \frac{[\gamma'(\theta)]^2}{I(\theta)}$ holds for every θ. The following counter example of Hodges & Lecam reveals that such an estimator does not, in general, exists.

If $\bar{X}$ is the sample mean based on n iid observations from a normal population with mean θ and variance unity, then $\sqrt{n}(\bar{X}-\theta) \xrightarrow{D} N(0, \sigma^2(\theta))$, where $\sigma(\theta) = 1$. Again we find that $I(\theta) = 1$ and hence $\sigma^2(\theta) = I(\theta)\ \forall\ \theta$. Thus according to Fisher, $\bar{X}$ is asymptotically efficient. Now define another estimator $T_n = a\bar{X}I(|\bar{X}| < n^{-1/4}) + \bar{X}I(|\bar{X}| > n^{-1/4})$, for some $a \in (0,1)$. Then it can be shown that $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \tau^2(\theta))$ with $\tau^2(\theta) = a^2$ or 1 as $\theta = 0$ or 1. Thus $\tau(0) < \sigma(0)$. Hence, asymptotically efficient estimators does not exist in the sense of Fisher.

3.2.3 Pitman's notion of asymptotic efficiency

Consider two sequence of estimators T_{1n} and T_{2n} for $\gamma(\theta)$ such that

$$\sqrt{n}(T_{1n} - \gamma(\theta)) \xrightarrow{D} N(0, \tau^2), \text{ and}$$
$$\sqrt{n}(T_{2n'} - \gamma(\theta)) \xrightarrow{D} N(0, \tau^2),$$

where $T_{2n'}$ is based on $n' = n'(n)$ observations. Then asymptotic relative efficiency(ARE) of T_{1n} with respect to T_{2n} is defined as $ARE(1|2) = \lim_{n\to\infty} \frac{n'(n)}{n}$ provided the limit exists and is independent of the subsequence $\{n'(n)\}$. Naturally, T_{1n} is better or worse asymptotically according as $ARE(1|2) > 1$ or $ARE(1|2) < 1$.
However, if

$$\sqrt{n}(T_{kn} - \gamma(\theta)) \xrightarrow{D} N(0, \sigma_k^2)$$

for $k = 1, 2$, then $ARE(1|2)$ can be expressed as σ_2^2/σ_1^2.

Let us consider an example to make the development clear. Consider comparison of sample mean and median as an estimator of a location parameter. To be specific, let X_i be iid observations from a DF $F(x-\theta)$, where $F(x) + F(-x) = 1 \forall x$ with finite variance σ^2. If $F'(x) = f(x)$ is non zero and continuous in an nbd of θ, then

$$\sqrt{n}(T_{1n} - \theta) \xrightarrow{D} N(0, \sigma^2), \text{ and}$$
$$\sqrt{n}(T_{2n} - \theta) \xrightarrow{D} N(0, \frac{1}{4f^2(\theta)}),$$

where T_{1n} is the sample mean and T_{2n} is the sample median. Then $ARE(2|1) = 4\sigma^2 f^2(\theta)$. For $N(\theta, 1)$, we find $ARE(2|1) = 2/\pi < 1$ whereas for $DE(\theta, 1)$ parent, $ARE(2|1) = 2 > 1$. Thus for double exponential parent, sample median is asymptotically more efficient than the sample mean.

Large-Sample Inference: Module 11[1]

Learn More

1. DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability, Springer.

2. Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, John Wiley, New York.

3. Shao, J. (2003). Mathematical Statistics, 2 nd ed. Springer, New York.

4. Van der Vaart, A. W. (2000). Asymptotic Statistics, Cambridge University Press.

5. Lehmann, E.L.(1999). Elements of Large-Sample Theory, Springer, New York.

6. Sen, P.K. and Singer, J.(1993). Large sample methods in statistics, Chapman & Hall, New York.

7. Ferguson, T.S.(1996). A course in large-sample theory. Chapman & Hall, New York.

8. Lecam, L. and Yang, G.L.(1990). Asymptotics in Statistics. Springer-Verlag, New York.

9. Rohatgi, V.K. and Saleh, A.K.(2002). An introduction to probability and statistics. Second Edition, John Wiley & Sons Inc., New York.

10. Tom M. Apostol(1974). Mathematical Analysis. Addison-Wesley Publishing Company, Inc..

[1]Co-Ordinator: Dr. Rahul Bhattacharya, Department of Statistics, Calcutta University

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Analysis

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Bayesian Analysis

Bayesian Analysis is a statistical methodology in which Bayes' Theorem is used to estimate the probability of a hypothesis as data is observed.

In this introductory module we discuss:

1. Frequentist and subjective interpretations of probability
2. Axioms of Probability
3. Independence and Conditional Probability
4. Bayes' Theorem and It's Application

In this introductory module we discuss:

1. Frequentist and subjective interpretations of probability

2. Axioms of Probability

3. Independence and Conditional Probability

4. Bayes' Theorem and It's Application

In this introductory module we discuss:

1. Frequentist and subjective interpretations of probability
2. Axioms of Probability
3. Independence and Conditional Probability
4. Bayes' Theorem and It's Application

Introduction

In this introductory module we discuss:

1. Frequentist and subjective interpretations of probability
2. Axioms of Probability
3. Independence and Conditional Probability
4. Bayes' Theorem and It's Application

The probability that some specific outcome is the relative frequency with which that outcome would be obtained if the process were repeated a large number of times under similar conditions.

Example

The probability of obtaining a head in a fair coin toss is 0.5 because the relative frequency of heads should be approximately 0.5 if we flip the coin many times.

The frequency interpretation

The probability that some specific outcome is the relative frequency with which that outcome would be obtained if the process were repeated a large number of times under similar conditions.

Example

The probability of obtaining a head in a fair coin toss is 0.5 because the relative frequency of heads should be approximately 0.5 if we flip the coin many times.

The frequency interpretation

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.
2. The sample is described by the sample mean and the sample variance.
3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

The frequency interpretation

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.

2. The sample is described by the sample mean and the sample variance.

3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

The frequency interpretation

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.

2. The sample is described by the sample mean and the sample variance.

3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

The frequency interpretation

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.
2. The sample is described by the sample mean and the sample variance.
3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.
2. The sample is described by the sample mean and the sample variance.
3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

The frequency interpretation

When we make statistical inferences from the frequentist perspective, we assume that our data is a sample from an entire population.

1. The population is described by the population mean and the population variance that are unknown.
2. The sample is described by the sample mean and the sample variance.
3. The sample mean and variance provide estimates about the mean and variance of the entire population.
4. Importantly, these estimates are known only with some uncertainty.

Our uncertainty about a statistic like the mean is summarized by its sampling distribution.

Sampling Distribution

The sampling distribution is a probability distribution of all possible values of a statistic of interest for samples of size N that could be formed for a given population.

- The observed sample mean is just one realization.

Problem

Beyond some text book cases finding the exact sampling distribution is difficult task.

Solution

Frequentist approach to the problem is to approximate the sampling distribution by known distribution like Gaussian or t distribution under the assumption like sample size N is large.

Critique 1

Needless to say, this is a theoretical construct since, with a large population, there will be billions of unique samples and it would be superior to simply sample the entire population.

The frequency interpretation

The sampling distribution

Problem

Beyond some text book cases finding the exact sampling distribution is difficult task.

Solution

Frequentist approach to the problem is to approximate the sampling distribution by known distribution like Gaussian or t distribution under the assumption like sample size N is large.

Critique 1

Needless to say, this is a theoretical construct since, with a large population, there will be billions of unique samples and it would be superior to simply sample the entire population.

Problem

Beyond some text book cases finding the exact sampling distribution is difficult task.

Solution

Frequentist approach to the problem is to approximate the sampling distribution by known distribution like Gaussian or t distribution under the assumption like sample size N is large.

Critique 1

Needless to say, this is a theoretical construct since, with a large population, there will be billions of unique samples and it would be superior to simply sample the entire population.

Problem

Beyond some text book cases finding the exact sampling distribution is difficult task.

Solution

Frequentist approach to the problem is to approximate the sampling distribution by known distribution like Gaussian or t distribution under the assumption like sample size N is large.

Critique 2

P-values refer to the proportion of hypothetical draws from the sampling distribution that are consistent with the null hypothesis. As p-values are based on the concept of a sampling distribution, do they make sense if our data contains the almost entire population?

Problem

Beyond some text book cases finding the exact sampling distribution is difficult task.

Solution

Frequentist approach to the problem is to approximate the sampling distribution by known distribution like Gaussian or t distribution under the assumption like sample size N is large.

Critique 2

P-values refer to the proportion of hypothetical draws from the sampling distribution that are consistent with the null hypothesis. As p-values are based on the concept of a sampling distribution, do they make sense if our data contains the almost entire population?

The Classical Interpretation of Probability

1. The classical interpretation is based on the concept of equally likely outcomes.

2. If the outcome of some process must be one of n different outcomes, and if these outcomes are equally likely to occur, then the probability of each outcome is $\frac{1}{n}$.

1. The classical interpretation is based on the concept of equally likely outcomes.

2. If the outcome of some process must be one of n different outcomes, and if these outcomes are equally likely to occur, then the probability of each outcome is $\frac{1}{n}$.

1. If we flip a fair coin, the probability of a head would be $\frac{1}{2}$ because head and tail are equally likely outcomes.

2. The classical approach offers an appealing summary of uncertainty in a one-shot situation.

1. If we flip a fair coin, the probability of a head would be $\frac{1}{2}$ because head and tail are equally likely outcomes.
2. The classical approach offers an appealing summary of uncertainty in a one-shot situation.

Problems with the Classical Interpretation

The drawback of the classical interpretation is that the concept of equally likely outcomes is itself probabilistic.

1. In a sense, this makes the classical definition of probability circular.

2. Furthermore, the concept begins to break down in contexts other than gambling when events are not equally likely.

Problems with the Classical Interpretation

The drawback of the classical interpretation is that the concept of equally likely outcomes is itself probabilistic.

1. In a sense, this makes the classical definition of probability circular.

2. Furthermore, the concept begins to break down in contexts other than gambling when events are not equally likely.

The classical response is...

1. Laplace's Rule of Insufficient Reason: in the absence of compelling evidence to the contrary, we should assume that events are equally likely.

2. This concept response is actually more useful to Bayesians when defending their priors.

The classical response is...

1. Laplace's Rule of Insufficient Reason: in the absence of compelling evidence to the contrary, we should assume that events are equally likely.

2. This concept response is actually more useful to Bayesians when defending their priors.

1. The probability that a person assigns to a possible outcome of some process represents his or her own judgment of the likelihood that the outcome will be obtained.

2. In contrast to the classical and frequentist interpretations of probability, this means that different individuals could have different probability judgments.

Example

If we flip a fair coin, the probability of a head could be $\frac{3}{4}$ because, for some reason, we think that *God wants it to be a head*.

The Subjective Interpretation of Probability

1. The probability that a person assigns to a possible outcome of some process represents his or her own judgment of the likelihood that the outcome will be obtained.

2. In contrast to the classical and frequentist interpretations of probability, this means that different individuals could have different probability judgments.

Example

If we flip a fair coin, the probability of a head could be $\frac{3}{4}$ because, for some reason, we think that *God wants it to be a head*.

The Subjective Interpretation of Probability

1. Is subjective probability theory really that *ad-hoc*?

2. Not Necessarily...Bayesian methodology elicit priors in a manner that ensures coherence.

1. Is subjective probability theory really that *ad-hoc*?

2. Not Necessarily...Bayesian methodology elicit priors in a manner that ensures coherence.

de Finetti Chapter 1

"This being granted, once an individual has evaluated the probabilities of certain events, two cases present themselves: either it is possible to bet with him in such a way as to be assured of winning, or else this probability does not exist. In the first case, one should say that the evaluation of probabilities given by this individual contains an incoherence, an intrinsic contradiction; in the other we say the individual is coherent. It is precisely this condition of coherence which constitutes the sole principle from which one can deduce the whole calculus of probability".

Extensions of de Finetti's axioms form the basis of subjective expected utility theory. Later chapters of the book introduce the concept of exchangeability, which is rather important to probability theory.

Axioms of Probability

A probability distribution on a sample space S is a specification of numbers $Pr(A_i)$ which satisfy:

Axiom 1

For any outcome A_i, $Pr(A_i) \geq 0$.

Axiom 2

$Pr(S) = 1$.

Axiom 3

For a sequence of disjoint events $A_1, A_2, \ldots$

$$Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Pr(A_i)$$

It turns out that each of these three axioms can be justified using the coherence criterion.

Axioms of Probability

A probability distribution on a sample space S is a specification of numbers $Pr(A_i)$ which satisfy:

Axiom 1

For any outcome A_i, $Pr(A_i) \geq 0$.

Axiom 2

$Pr(S) = 1$.

Axiom 3

For a sequence of disjoint events $A_1, A_2, ...$

$$Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Pr(A_i)$$

It turns out that each of these three axioms can be justified using the coherence criterion.

A probability distribution on a sample space S is a specification of numbers $Pr(A_i)$ which satisfy:

Axiom 1

For any outcome A_i, $Pr(A_i) \geq 0$.

Axiom 2

$Pr(S) = 1$.

Axiom 3

For a sequence of disjoint events $A_1, A_2, ...$

$$Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Pr(A_i)$$

It turns out that each of these three axioms can be justified using the coherence criterion.

Theorems of Probability

1. $P(\Phi) = 0$

2. For any finite sequence of disjoint events $\{A_1, A_2, ..., A_n\}$

$$Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$$

3. For any event A, $Pr(A^c) = 1 - Pr(A)$

4. For any event A, $0 \leq Pr(A) \leq 1$

5. For any two events A and B,
$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

1. $P(\Phi) = 0$

2. For any finite sequence of disjoint events $\{A_1, A_2, ..., A_n\}$

$$Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$$

3. For any event A, $Pr(A^c) = 1 - Pr(A)$

4. For any event A, $0 \leq Pr(A) \leq 1$

5. For any two events A and B,
$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

1. $P(\Phi) = 0$

2. For any finite sequence of disjoint events $\{A_1, A_2, ..., A_n\}$

$$Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$$

3. For any event A, $Pr(A^c) = 1 - Pr(A)$

4. For any event A, $0 \leq Pr(A) \leq 1$

5. For any two events A and B,
$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Theorems of Probability

1. $P(\Phi) = 0$

2. For any finite sequence of disjoint events $\{A_1, A_2, ..., A_n\}$

$$Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$$

3. For any event A, $Pr(A^c) = 1 - Pr(A)$

4. For any event A, $0 \leq Pr(A) \leq 1$

5. For any two events A and B,
$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

1. $P(\Phi) = 0$

2. For any finite sequence of disjoint events $\{A_1, A_2, ..., A_n\}$

$$Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$$

3. For any event A, $Pr(A^c) = 1 - Pr(A)$

4. For any event A, $0 \leq Pr(A) \leq 1$

5. For any two events A and B,
$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Independence Events

Idea of Independence

Two events A and B are independent if the occurrence or non-occurrence of one of the events has no influence on the occurrence or non-occurrence of the other event.

Independence Mathematical Definition

Two events A and B are independent if

$$Pr(A \cap B) = Pr(A)Pr(B)$$

Independence Events

Idea of Independence

Two events A and B are independent if the occurrence or non-occurrence of one of the events has no influence on the occurrence or non-occurrence of the other event.

Independence Mathematical Definition

Two events A and B are independent if

$$Pr(A \cap B) = Pr(A)Pr(B)$$

Are 'Smoking' and 'Lung Cancer' independent?

Suppose

1. $Pr(Smoking) = 0.4$,
2. $Pr(Lung\ Cancer) = 0.5$,
3. $Pr(Lung\ Cancer\ \cap\ Smoking) = 0.35$

Clearly, $Pr(Lung\ Cancer) \times Pr(Smoking) = 0.4 \times 0.5 = 0.2 \neq 0.35 = Pr(Lung\ Cancer\ \cap\ Smoking)$

Are 'Smoking' and 'Lung Cancer' independent?

Suppose

1. $Pr(Smoking) = 0.4$,
2. $Pr(Lung\ Cancer) = 0.5$,
3. $Pr(Lung\ Cancer\ \cap\ Smoking) = 0.35$

Clearly, $Pr(Lung\ Cancer) \times Pr(Smoking) = 0.4 \times 0.5 = 0.2 \neq 0.35 = Pr(Lung\ Cancer\ \cap\ Smoking)$

Are 'Smoking' and 'Lung Cancer' independent?

Suppose

1. $Pr(Smoking) = 0.4$,
2. $Pr(Lung\ Cancer) = 0.5$,
3. $Pr(Lung\ Cancer\ \cap\ Smoking) = 0.35$

Clearly, $Pr(Lung\ Cancer) \times Pr(Smoking) = 0.4 \times 0.5 = 0.2 \neq 0.35 = Pr(Lung\ Cancer\ \cap\ Smoking)$

Are 'Smoking' and 'Lung Cancer' independent?

Suppose

1. $Pr(Smoking) = 0.4$,

2. $Pr(Lung\ Cancer) = 0.5$,

3. $Pr(Lung\ Cancer\ \cap\ Smoking) = 0.35$

Clearly, $Pr(Lung\ Cancer) \times Pr(Smoking) = 0.4 \times 0.5 = 0.2 \neq 0.35 = Pr(Lung\ Cancer\ \cap\ Smoking)$

Conditional probabilities allow us to understand how the probability of an event A changes after it has been learned that some other event B has occurred.

- The key concept for thinking about conditional probabilities is that the occurrence of B reshapes the sample space for subsequent events.

- That is, we begin with a sample space S

- A and $B \in S$

- The conditional probability of A given that B looks just at the subset of the sample space for B.

Conditional probabilities allow us to understand how the probability of an event A changes after it has been learned that some other event B has occurred.

- The key concept for thinking about conditional probabilities is that the occurrence of B reshapes the sample space for subsequent events.
- That is, we begin with a sample space S
- A and $B \in S$
- The conditional probability of A given that B looks just at the subset of the sample space for B.

Conditional probabilities allow us to understand how the probability of an event A changes after it has been learned that some other event B has occurred.

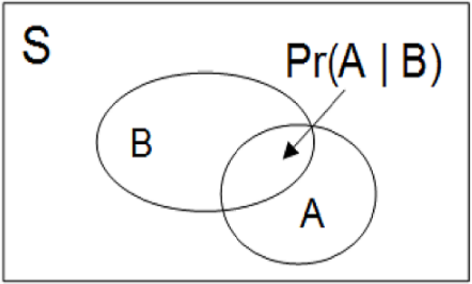
- The key concept for thinking about conditional probabilities is that the occurrence of B reshapes the sample space for subsequent events.
- That is, we begin with a sample space S
- A and $B \in S$
- The conditional probability of A given that B looks just at the subset of the sample space for B.

Conditional Probability

Conditional probabilities allow us to understand how the probability of an event A changes after it has been learned that some other event B has occurred.

- The key concept for thinking about conditional probabilities is that the occurrence of B reshapes the sample space for subsequent events.
- That is, we begin with a sample space S
- A and $B \in S$
- The conditional probability of A given that B looks just at the subset of the sample space for B.

Conditional Probability



1. The conditional probability of A given B is denoted $Pr(A|B)$.

2. Importantly, according to Bayesian orthodoxy, all probability distributions are implicitly or explicitly conditioned on the model.

3. By definition: If A and B are two events such that $Pr(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

1. The conditional probability of A given B is denoted $Pr(A|B)$.
2. Importantly, according to Bayesian orthodoxy, all probability distributions are implicitly or explicitly conditioned on the model.
3. By definition: If A and B are two events such that $Pr(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

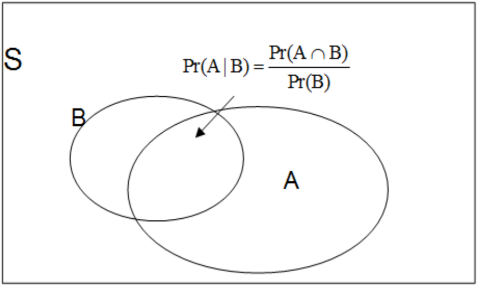
1. The conditional probability of A given B is denoted $Pr(A|B)$.

2. Importantly, according to Bayesian orthodoxy, all probability distributions are implicitly or explicitly conditioned on the model.

3. By definition: If A and B are two events such that $Pr(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability



Example

1. What is the $Pr(Lung\ Cancer|Smoking)$?
2. $Pr(Lung\ Cancer \cap Smoking) = 0.35$
3. $Pr(Smoking) = .4$
4. Thus, $Pr(Lung\ Cancer|Smoking) = .35/.4 = .875$

ePathshala पाठशाला MHRD Govt. of India

Example

1. What is the $Pr(Lung\ Cancer|Smoking)$?

2. $Pr(Lung\ Cancer \cap Smoking) = 0.35$

3. $Pr(Smoking) = .4$

4. Thus, $Pr(Lung\ Cancer|Smoking) = .35/.4 = .875$

Conditional Probability

Example

1. What is the $Pr(Lung\ Cancer|Smoking)$?

2. $Pr(Lung\ Cancer \cap Smoking) = 0.35$

3. $Pr(Smoking) = .4$

4. Thus, $Pr(Lung\ Cancer|Smoking) = .35/.4 = .875$

Example

1. What is the $Pr(Lung\ Cancer|Smoking)$?
2. $Pr(Lung\ Cancer \cap Smoking) = 0.35$
3. $Pr(Smoking) = .4$
4. Thus, $Pr(Lung\ Cancer|Smoking) = .35/.4 = .875$

1. **The Conditional Probability for Independent Events**: If A and B are independent then $P(A|B) = P(A)$

2. **The Multiplication Rule for Conditional Probabilities**: In an experiment involving two non-independent events A and B, the probability that both A and B occurs can be found in the following two ways:

$$Pr(A \cap B) = Pr(B)Pr(A|B)$$

or

$$Pr(A \cap B) = Pr(A)Pr(B|A)$$

3. The set of events $\{A_1, ..., A_n\}$ are partition of sample space S, where $\cup_{i=1}^{n} = S$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

1. **The Conditional Probability for Independent Events**: If A and B are independent then $P(A|B) = P(A)$

2. **The Multiplication Rule for Conditional Probabilities**: In an experiment involving two non-independent events A and B, the probability that both A and B occurs can be found in the following two ways:

$$Pr(A \cap B) = Pr(B)Pr(A|B)$$

or

$$Pr(A \cap B) = Pr(A)Pr(B|A)$$

3. The set of events $\{A_1, ..., A_n\}$ are partition of sample space S, where $\cup_{i=1}^{n} = S$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Properties of Conditional Probability

1. **The Conditional Probability for Independent Events**: If A and B are independent then $P(A|B) = P(A)$

2. **The Multiplication Rule for Conditional Probabilities**: In an experiment involving two non-independent events A and B, the probability that both A and B occurs can be found in the following two ways:

$$Pr(A \cap B) = Pr(B)Pr(A|B)$$

   or

$$Pr(A \cap B) = Pr(A)Pr(B|A)$$

3. The set of events $\{A_1, ..., A_n\}$ are partition of sample space S, where $\cup_{i=1}^{n} = S$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes' Theorem

Let events $A_1, ..., A_k$ form a partition of the space S such that $Pr(A_j) > 0$ for all j and let B be any event such that Pr(B) > 0. Then for i = 1,..,k:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

1. Bayes' Theorem is just a simple rule for computing the conditional probability of events A_i given B from the conditional probability of B given each event A_i and the unconditional probability of each A_i

2. $P(A_i)$ is the prior distribution of A_i.

3. $P(B|A_i)$ is the the conditional probability of B given A_i

4. $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ is the normalizing constant

5. $P(A_i|B)$ is the posterior probability of A_i given B

Bayes' Theorem

1. Bayes' Theorem is just a simple rule for computing the conditional probability of events A_i given B from the conditional probability of B given each event A_i and the unconditional probability of each A_i

2. $P(A_i)$ is the prior distribution of A_i.

3. $P(B|A_i)$ is the the conditional probability of B given A_i

4. $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ is the normalizing constant

5. $P(A_i|B)$ is the posterior probability of A_i given B

1. Bayes' Theorem is just a simple rule for computing the conditional probability of events A_i given B from the conditional probability of B given each event A_i and the unconditional probability of each A_i

2. $P(A_i)$ is the prior distribution of A_i.

3. $P(B|A_i)$ is the the conditional probability of B given A_i

4. $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ is the normalizing constant

5. $P(A_i|B)$ is the posterior probability of A_i given B

1. Bayes' Theorem is just a simple rule for computing the conditional probability of events A_i given B from the conditional probability of B given each event A_i and the unconditional probability of each A_i
2. $P(A_i)$ is the prior distribution of A_i.
3. $P(B|A_i)$ is the the conditional probability of B given A_i
4. $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ is the normalizing constant
5. $P(A_i|B)$ is the posterior probability of A_i given B

1. Bayes' Theorem is just a simple rule for computing the conditional probability of events A_i given B from the conditional probability of B given each event A_i and the unconditional probability of each A_i

2. $P(A_i)$ is the prior distribution of A_i.

3. $P(B|A_i)$ is the the conditional probability of B given A_i

4. $\sum_{i=1}^{n} P(B|A_i)P(A_i)$ is the normalizing constant

5. $P(A_i|B)$ is the posterior probability of A_i given B

e-Pathshala पाठशाला

A Gateway to all Postgraduate Courses

MHRD Govt. of India

सत्यमेव जयते

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference

Module: Bayesian Analysis of One Parameter Model

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

1. The binomial distribution-uniform prior integration tricks
2. Posterior Interpretation
3. Binomial distribution-beta prior
4. Conjugate priors and sufficient statistics

Modeling the Unknown Quantities

From the Bayesian perspective, there are known and unknown quantities.

- The known quantity is the data, denoted D.
- The unknown quantities are the parameters (e.g. mean, variance, missing data), denoted θ.

To make inferences about the unknown quantities, we stipulate a joint probability function that describes how we believe these quantities behave in conjunction, $p(\theta, D)$.

- Using Bayes' Rule, this joint probability function can be rearranged to make inference about θ

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{L(\theta|D)p(\theta)}{\int L(\theta|D)p(\theta)d\theta} \end{aligned}$$

- $L(\theta|D)$ is the likelihood function of θ

- $\int L(\theta|D)p(\theta)d\theta$ is the normalizing constant or prior predictive distribution

- It is the normalizing constant because it ensures that the posterior distribution of θ integrates to one.

- Using Bayes' Rule, this joint probability function can be rearranged to make inference about θ

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{L(\theta|D)p(\theta)}{\int L(\theta|D)p(\theta)d\theta} \end{aligned}$$

- $L(\theta|D)$ is the likelihood function of θ
- $\int L(\theta|D)p(\theta)d\theta$ is the normalizing constant or prior predictive distribution
- It is the normalizing constant because it ensures that the posterior distribution of θ integrates to one.

Bayesian Setup

- Using Bayes' Rule, this joint probability function can be rearranged to make inference about θ

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{L(\theta|D)p(\theta)}{\int L(\theta|D)p(\theta)d\theta} \end{aligned}$$

- $L(\theta|D)$ is the likelihood function of θ

- $\int L(\theta|D)p(\theta)d\theta$ is the normalizing constant or prior predictive distribution

- It is the normalizing constant because it ensures that the posterior distribution of θ integrates to one.

- It is the prior predictive distribution because it is not conditional on a previous observation of the data-generating process (prior) and because it is the distribution of an observable quantity (predictive).

Popular Presentation

The posterior distribution often presented as

$$p(\theta|D) \propto L(\theta|D)p(\theta)$$

i.e., posterior $\propto$ likelihood $\times$ prior

- Why are we allowed to do this?

- Why might not be as useful?

- It is the prior predictive distribution because it is not conditional on a previous observation of the data-generating process (prior) and because it is the distribution of an observable quantity (predictive).

Popular Presentation

The posterior distribution often presented as

$$p(\theta|D) \propto L(\theta|D)p(\theta)$$

i.e., posterior $\propto$ likelihood $\times$ prior

- Why are we allowed to do this?
- Why might not be as useful?

- It is the prior predictive distribution because it is not conditional on a previous observation of the data-generating process (prior) and because it is the distribution of an observable quantity (predictive).

Popular Presentation

The posterior distribution often presented as

$$p(\theta|D) \propto L(\theta|D)p(\theta)$$

i.e., posterior $\propto$ likelihood $\times$ prior

- Why are we allowed to do this?

- Why might not be as useful?

Bayesian Analysis: Binomial Distribution

- Suppose X_1, X_2,...,X_n are independent random draws from same Bernoulli distribution with parameter π (unknown).
- Thus $X_i \sim Bernoulli(\pi)$ for $i \in \{1, 2, ..., n\}$ or equivalently $Y = \sum_{i=1}^{n} X_i \sim Bin(n, \pi)$.
- The joint distribution of Y and π is the product of the conditional distribution of Y and the prior distribution π.
- What distribution might be a reasonable choice for the prior distribution of π? Why?

- Suppose X_1, X_2,...,X_n are independent random draws from same Bernoulli distribution with parameter π (unknown).
- Thus $X_i \sim Bernoulli(\pi)$ for $i \in \{1, 2, ..., n\}$ or equivalently $Y = \sum_{i=1}^{n} X_i \sim Bin(n, \pi)$.
- The joint distribution of Y and π is the product of the conditional distribution of Y and the prior distribution π.
- What distribution might be a reasonable choice for the prior distribution of π? Why?

- Suppose X_1, X_2,...,X_n are independent random draws from same Bernoulli distribution with parameter π (unknown).
- Thus $X_i \sim Bernoulli(\pi)$ for $i \in \{1, 2, ..., n\}$ or equivalently $Y = \sum_{i=1}^{n} X_i \sim Bin(n, \pi)$.
- The joint distribution of Y and π is the product of the conditional distribution of Y and the prior distribution π.
- What distribution might be a reasonable choice for the prior distribution of π? Why?

- Suppose X_1, X_2,...,X_n are independent random draws from same Bernoulli distribution with parameter π (unknown).
- Thus $X_i \sim Bernoulli(\pi)$ for $i \in \{1, 2, ..., n\}$ or equivalently $Y = \sum_{i=1}^{n} X_i \sim Bin(n, \pi)$.
- The joint distribution of Y and π is the product of the conditional distribution of Y and the prior distribution π.
- What distribution might be a reasonable choice for the prior distribution of π? Why?

- If $Y\ Bin(n,\pi)$, a reasonable prior distribution for p must be bounded between zero and one.

  One option is the uniform distribution $\pi \sim Unif(0,1)$.

-
$$p(\pi|Y) \propto^n C_y \pi^y (1-\pi)^{n-y} \times 1$$

- As it happens, this is a proper posterior density function.

- How can you tell?

- If $Y\ Bin(n, \pi)$, a reasonable prior distribution for p must be bounded between zero and one.

  One option is the uniform distribution $\pi \sim Unif(0, 1)$.

- $$p(\pi|Y) \propto^n C_y \pi^y (1 - \pi)^{n-y} \times 1$$

- As it happens, this is a proper posterior density function.

- How can you tell?

- If $Y\ Bin(n, \pi)$, a reasonable prior distribution for p must be bounded between zero and one.
  One option is the uniform distribution $\pi \sim Unif(0, 1)$.
- $$p(\pi|Y) \propto^n C_y \pi^y (1 - \pi)^{n-y} \times 1$$
- As it happens, this is a proper posterior density function.
- How can you tell?

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0, 1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^{n}C_{y}\pi^{y}(1-\pi)^{n-y} \times 1 \\ &\propto \pi^{y}(1-\pi)^{n-y} \end{aligned}$$

- You cannot just call the posterior a binomial distribution because you are conditioning on Y and π is a random variable, not the other way around.

- The pdf of beta distribution which is known to be proper is:

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0, 1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^nC_y \pi^y (1-\pi)^{n-y} \times 1 \\ &\propto \pi^y (1-\pi)^{n-y} \end{aligned}$$

- You cannot just call the posterior a binomial distribution because you are conditioning on Y and π is a random variable, not the other way around.

- The pdf of beta distribution which is known to be proper is:

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0, 1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^{n}C_{y}\pi^{y}(1-\pi)^{n-y} \times 1 \\ &\propto \pi^{y}(1-\pi)^{n-y} \end{aligned}$$

- You cannot just call the posterior a binomial distribution because you are conditioning on Y and π is a random variable, not the other way around.

- The pdf of beta distribution which is known to be proper is:

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

Bayesian Analysis: Binomial Distribution

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0,1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^nC_y \pi^y (1-\pi)^{n-y} \times 1 \\ &\propto \pi^y (1-\pi)^{n-y} \end{aligned}$$

- The pdf of beta distribution which is known to be proper is:
$$Beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

- Let $x = \pi$, $\alpha = Y + 1$ and $\beta = n - Y + 1$

- Thus $p(\pi|Y = y) \sim Beta(y+1, n-y+1) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \pi^{(y+1)-1}(1-\pi)^{(n-y+1)-1}$

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0, 1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^nC_y \pi^y (1-\pi)^{n-y} \times 1 \\ &\propto \pi^y (1-\pi)^{n-y} \end{aligned}$$

- The pdf of beta distribution which is known to be proper is:

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

- Let $x = \pi$, $\alpha = Y + 1$ and $\beta = n - Y + 1$

- Thus $p(\pi|Y = y) \sim Beta(y + 1, n - y + 1) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \pi^{(y+1)-1}(1-\pi)^{(n-y+1)-1}$

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0,1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^nC_y \pi^y (1-\pi)^{n-y} \times 1 \\ &\propto \pi^y (1-\pi)^{n-y} \end{aligned}$$

- The pdf of beta distribution which is known to be proper is:
$$Beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
where $0 < x < 1$ and $\alpha > 0$, $\beta > 0$
Note that $\Gamma(k)$ is the Gamma function.

- Let $x = \pi$, $\alpha = Y + 1$ and $\beta = n - Y + 1$

- Thus $p(\pi|Y = y) \sim Beta(y+1, n-y+1) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \pi^{(y+1)-1}(1-\pi)^{(n-y+1)-1}$

Let $Y \sim Bin(n, \pi)$ and $\pi \sim unif(0, 1)$

- $$\begin{aligned} p(\pi|Y) &\propto {}^{n}C_y \pi^y (1-\pi)^{n-y} \times 1 \\ &\propto \pi^y (1-\pi)^{n-y} \end{aligned}$$

  Note that $\Gamma(k)$ is the Gamma function.

- Let $x = \pi$, $\alpha = Y + 1$ and $\beta = n - Y + 1$
- Thus $p(\pi|Y = y) \sim Beta(y + 1, n - y + 1) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \pi^{(y+1)-1} (1-\pi)^{(n-y+1)-1}$
- Note that $\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}$ is the normalizing constant to transform $\pi^y (1-\pi)^{n-y}$

Description

A researcher examined the level of consensus denoted π among $n = 24$ Guatemalan women about whether or not polio (as well as other diseases) was thought to be contagious. In this case, 17 women said polio was contagious.

- Let $X_i = 1$ if respondent i thought polio was contagious and $X_i = 0$ otherwise

- Let $\sum_{i=1}^{2} 4X_i = Y \sim Bin(24, \pi)$ and $\pi \sim Unif(0, 1)$ []
- $p(\pi|Y, n) \sim Beta(y + 1, n - y + 1)$

- Substitute $n = 24$ and $Y = 17$ we have the posterior distribution as

$$p(\pi|Y, n) = Beta(18, 8)$$

Description

A researcher examined the level of consensus denoted π among $n = 24$ Guatemalan women about whether or not polio (as well as other diseases) was thought to be contagious. In this case, 17 women said polio was contagious.

- Let $X_i = 1$ if respondent i thought polio was contagious and $X_i = 0$ otherwise

- Let $\sum_{i=1}^{2} 4X_i = Y \sim Bin(24, \pi)$ and $\pi \sim Unif(0, 1)$ []
- $p(\pi|Y, n) \sim Beta(y + 1, n - y + 1)$

- Substitute $n = 24$ and $Y = 17$ we have the posterior distribution as

$$p(\pi|Y, n) = Beta(18, 8)$$

Description

A researcher examined the level of consensus denoted π among $n = 24$ Guatemalan women about whether or not polio (as well as other diseases) was thought to be contagious. In this case, 17 women said polio was contagious.

- Let $X_i = 1$ if respondent i thought polio was contagious and $X_i = 0$ otherwise

- Let $\sum_{i=1}^{2} 4X_i = Y \sim Bin(24, \pi)$ and $\pi \sim Unif(0, 1)$ []
- $p(\pi|Y, n) \sim Beta(y + 1, n - y + 1)$

- Substitute $n = 24$ and $Y = 17$ we have the posterior distribution as

$$p(\pi|Y, n) = Beta(18, 8)$$

Description

A researcher examined the level of consensus denoted π among $n = 24$ Guatemalan women about whether or not polio (as well as other diseases) was thought to be contagious. In this case, 17 women said polio was contagious.

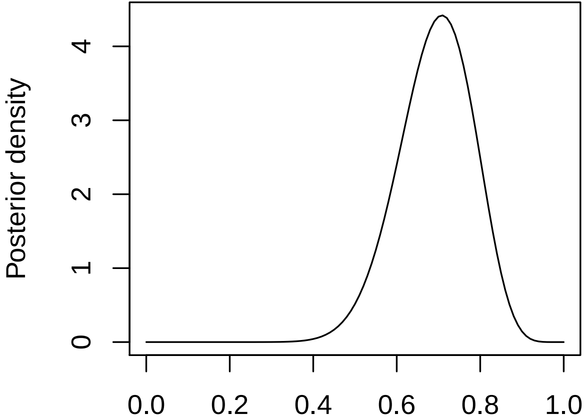
- Let $X_i = 1$ if respondent i thought polio was contagious and $X_i = 0$ otherwise

- Let $\sum_{i=1}^{2} 4X_i = Y \sim Bin(24, \pi)$ and $\pi \sim Unif(0, 1)$ []
- $p(\pi|Y, n) \sim Beta(y + 1, n - y + 1)$

- Substitute $n = 24$ and $Y = 17$ we have the posterior distribution as

$$p(\pi|Y, n) = Beta(18, 8)$$

Posterior density plot of π:



The full posterior contains too much information, especially in multi-parameter models. So, we use summary statistics (e.g. mean, var, HDR).

- Methods for generating summary stats:
  - Analytical Solutions: use the well-known analytic solutions for the mean, variance, etc. of the various posterior distribution.
  - Numerical Solutions: use a random number generator to draw a large number of values from the posterior distribution, then compute summary stats from those random draws.

The full posterior contains too much information, especially in multi-parameter models. So, we use summary statistics (e.g. mean, var, HDR).

- Methods for generating summary stats:
  - Analytical Solutions: use the well-known analytic solutions for the mean, variance, etc. of the various posterior distribution.
  - Numerical Solutions: use a random number generator to draw a large number of values from the posterior distribution, then compute summary stats from those random draws.

The full posterior contains too much information, especially in multi-parameter models. So, we use summary statistics (e.g. mean, var, HDR).

- Methods for generating summary stats:
  - Analytical Solutions: use the well-known analytic solutions for the mean, variance, etc. of the various posterior distribution.
  - Numerical Solutions: use a random number generator to draw a large number of values from the posterior distribution, then compute summary stats from those random draws.

- Analytic summaries are based on standard results from probability theory
- Continuing our example, $p(\pi|Y,n) = Beta(18,8)$
- If $\theta \sim Beta(\alpha,\beta)$

$$\begin{array}{rcl} \mathrm{E}(\theta) = \frac{\alpha}{\alpha+\beta} & & \mathrm{E}(\pi|Y,n) = \frac{18}{18+8} = 0.69 \\ \mathrm{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} & & \mathrm{Var}(\pi|Y,n) = \frac{18\ 8}{(18+8)^2(18+8+1)} = 0.01 \\ \mathrm{Mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2} & & \mathrm{Mode}(\pi|Y,n) = \frac{18-1}{18+8-2} = 0.71 \end{array}$$

Analytic Summaries of the Posterior

- Analytic summaries are based on standard results from probability theory

- Continuing our example, $p(\pi|Y,n) = Beta(18,8)$

- If $\theta \sim Beta(\alpha,\beta)$

$$
\begin{aligned}
\mathsf{E}(\theta) &= \tfrac{\alpha}{\alpha+\beta} & \mathsf{E}(\pi|Y,n) &= \tfrac{18}{18+8} = 0.69 \\
\mathsf{Var}(\theta) &= \tfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} & \mathsf{Var}(\pi|Y,n) &= \tfrac{18 \;\; 8}{(18+8)^2(18+8+1)} = 0.01 \\
\mathsf{Mode}(\theta) &= \tfrac{\alpha-1}{\alpha+\beta-2} & \mathsf{Mode}(\pi|Y,n) &= \tfrac{18-1}{18+8-2} = 0.71
\end{aligned}
$$

Numerical Summaries of the Posterior

- To create numerical summaries from the posterior, you need a random number generator.

- To summarize $p(\pi|Y,n) \sim Beta(18,8)$
  - Draw large number of random samples from $Beta(18,8)$ distribution.
  - Calculate the sample statistics from that set of random samples.

Numerical Summaries of the Posterior

- To create numerical summaries from the posterior, you need a random number generator.
- To summarize $p(\pi|Y, n) \sim Beta(18, 8)$
  - Draw large number of random samples from $Beta(18, 8)$ distribution.
  - Calculate the sample statistics from that set of random samples.

Numerical Summaries of the Posterior

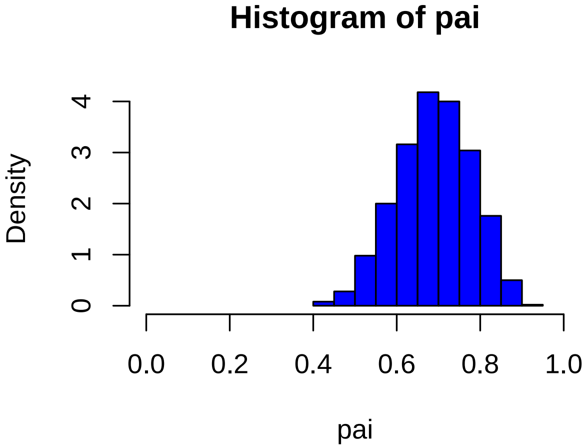
- To create numerical summaries from the posterior, you need a random number generator.

- To summarize $p(\pi|Y, n) \sim Beta(18, 8)$
  - Draw large number of random samples from $Beta(18, 8)$ distribution.

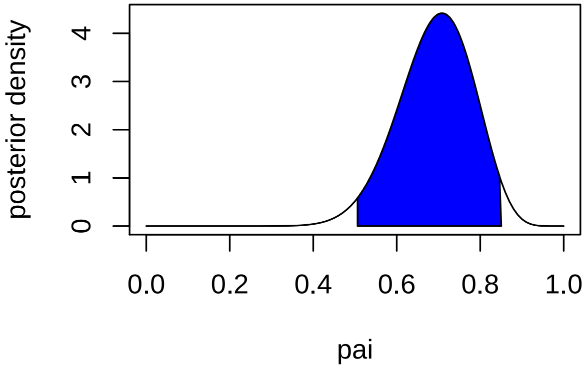
  - Calculate the sample statistics from that set of random samples.

Numerical Summaries of the Posterior

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4045	0.6311	0.6916	0.6903	0.7539	0.9065



Highest Density Regions (HDR's) are intervals containing a specified posterior probability. The figure below plots the 95% highest posterior density region.



Bayesian credible interval

The Bayesian credible interval is the probability that a true value of θ lies in the interval. Technically,

$$P(\theta \in Interval|Data) = \alpha$$

Note that here probability statement is direct.

Frequentist Confidence interval

The Frequentist Confidence interval is the region of sampling distribution for θ such that given the observed data one would expect (1-α) percent of the future estimates of θ to be outside that interval. Note that here understanding of probability is implicit. It is not a direct probability statement.

Confidence Intervals vs. Bayesian Credible Intervals

Bayesian credible interval

The Bayesian credible interval is the probability that a true value of θ lies in the interval. Technically,

$$P(\theta \in Interval|Data) = \alpha$$

Note that here probability statement is direct.

Frequentist Confidence interval

The Frequentist Confidence interval is the region of sampling distribution for θ such that given the observed data one would expect (1-α) percent of the future estimates of θ to be outside that interval. Note that here understanding of probability is implicit. It is not a direct probability statement.

Confidence Intervals vs. Bayesian Credible Intervals

- But often the results appear similar.
- If Bayesians use "non-informative priors" and there is a large number of observations, often several dozen will do, HDRs and frequentist confidence intervals will coincide numerically.
- We will talk more about this when we cover the great p-value debate, but this is only a coincidence.
- The interpretation of the two quantities are entirely different.

Confidence Intervals vs. Bayesian Credible Intervals

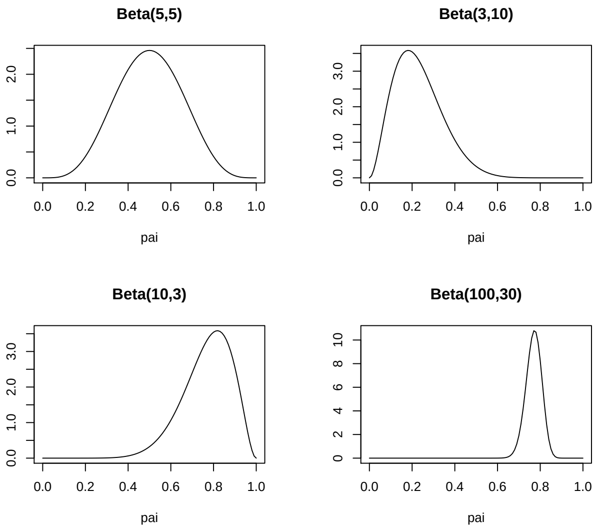
- But often the results appear similar.
- If Bayesians use "non-informative priors" and there is a large number of observations, often several dozen will do, HDRs and frequentist confidence intervals will coincide numerically.
- We will talk more about this when we cover the great p-value debate, but this is only a coincidence.
- The interpretation of the two quantities are entirely different.

Confidence Intervals vs. Bayesian Credible Intervals

- But often the results appear similar.
- If Bayesians use "non-informative priors" and there is a large number of observations, often several dozen will do, HDRs and frequentist confidence intervals will coincide numerically.
- We will talk more about this when we cover the great p-value debate, but this is only a coincidence.
- The interpretation of the two quantities are entirely different.

- If $Y \sim Bin(n, \pi)$, the uniform prior is just one of an infinite number of possible prior distributions.
- What other distributions could we use?
- A reasonable alternative to the unif(0,1) distribution is the beta distribution.
- Can you show that Beta(1,1) is a uniform(0,1) distribution?

Prior Consequences
Plots of 4 Different Beta Distributions



Binomial Distribution with Beta Prior

- If $Y \sim Bin(n, \pi)$ and $\pi \sim Beta(\alpha, \beta)$

- The posterior distribution:
$$p(\pi|Y,n) = \frac{{}^{n}C_{y}\pi^{y}(1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{\int_0^1 {}^{n}C_{y}\pi^{y}(1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi}$$

Note

This is a very complicated integral in the denominator. Though this particular integral can be solved; but we will pretend that it is difficult integral and we shall use a standard trick in the Bayesian toolbox to solve this problem.

- If $Y \sim Bin(n, \pi)$ and $\pi \sim Beta(\alpha, \beta)$

- The posterior distribution:

$$\text{p}(\pi|Y,n) = \frac{{}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}}{\int_0^1 {}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi}$$

Note

This is a very complicated integral in the denominator. Though this particular integral can be solved; but we will pretend that it is difficult integral and we shall use a standard trick in the Bayesian toolbox to solve this problem.

Trick:

- Find some multiplicative constant c such that $f(y) * c = 1$.
  i.e., try to transform $f(y)$ into a well-known pdf.

- Multiply c and c^{-1}

- Since $c * f(y) = 1$, the original numerator multiplied by c^{-1} is the posterior distribution.

Trick:

- Find some multiplicative constant c such that $f(y) * c = 1$. i.e., try to transform $f(y)$ into a well-known pdf.
- Multiply c and c^{-1}
- Since $c * f(y) = 1$, the original numerator multiplied by c^{-1} is the posterior distribution.

Trick:

- Find some multiplicative constant c such that $f(y) * c = 1$. i.e., try to transform $f(y)$ into a well-known pdf.
- Multiply c and c^{-1}
- Since $c * f(y) = 1$, the original numerator multiplied by c^{-1} is the posterior distribution.

The posterior predictive distribution

$$f(y) = \int_0^1 {}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1} d\pi$$

- It can be expressed as:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{y+\alpha-1} (1-\pi)^{n+\beta-y-1} d\pi$$

- $\int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$ is the kernel of the beta distribution

-

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

This is called beta-binomial distribution

The posterior predictive distribution

$$f(y) = \int_0^1 {}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi$$

- It can be expressed as:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$$

- $\int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$ is the kernel of the beta distribution

-

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

This is called beta-binomial distribution

$$f(y) = \int_0^1 {}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi$$

- It can be expressed as:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$$

- $\int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$ is the kernel of the beta distribution

-

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

This is called beta-binomial distribution

$$f(y) = \int_0^1 {}^nC_y \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi$$

- It can be expressed as:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$$

- $\int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1} d\pi$ is the kernel of the beta distribution

- $$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

This is called beta-binomial distribution

- As the marginal distribution is:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

- The posterior distribution is :

$$f(\pi|Y) = \frac{f(y|\pi)f(\pi)}{f(y)}$$

$$f(\pi|Y) = \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^y(1-\pi)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{f(y)}$$

- As the marginal distribution is:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

- The posterior distribution is :

$$f(\pi|Y) = \frac{f(y|\pi)f(\pi)}{f(y)}$$

$$f(\pi|Y) = \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^{y}(1-\pi)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{f(y)}$$

e-pathshala पाठशाला MHRD Govt. of India

- As the marginal distribution is:

$$f(y) = \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+n+\beta)}$$

- The posterior distribution is :

$$f(\pi|Y) = \frac{f(y|\pi)f(\pi)}{f(y)}$$

$$f(\pi|Y) = \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^y(1-\pi)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{f(y)}$$

- The posterior distribution is :

$$f(\pi|y) = \frac{f(y|\pi)f(\pi)}{f(y)}$$

$$f(\pi|y) = \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^y(1-\pi)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{f(y)}$$

- Simplifying the above expression:

$$f(\pi|y) = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1}$$

- This is $Beta(y+\alpha, n-y+\beta)$ distribution.

- The posterior distribution is :

$$f(\pi|y) = \frac{f(y|\pi)f(\pi)}{f(y)}$$

$$f(\pi|y) = \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^y(1-\pi)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{f(y)}$$

- Simplifying the above expression:

$$f(\pi|y) = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\pi^{y+\alpha-1}(1-\pi)^{n+\beta-y-1}$$

- This is $Beta(y+\alpha, n-y+\beta)$ distribution.

Note

You can see posterior distribution has the same distribution as prior distribution updated by new data. In general, when this happens we say the prior is conjugate.

Application

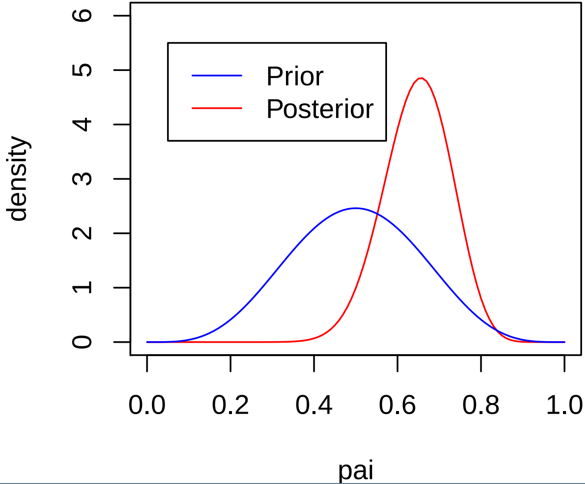
Lets continue to the previous example. You remeber 17 of 24 women say polio is contagious (so $y = 17$ and $n = 24$ where y is a realization from binomial) and you use $Beta(5, 5)$ prior; the posterior distribution is $Beta(17 + 5, 24 - 17 + 5) = Beta(22, 12)$

Note

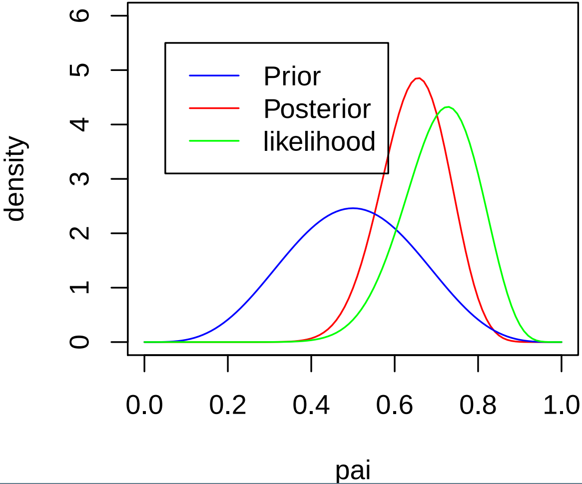
You can see posterior distribution has the same distribution as prior distribution updated by new data. In general, when this happens we say the prior is conjugate.

Application

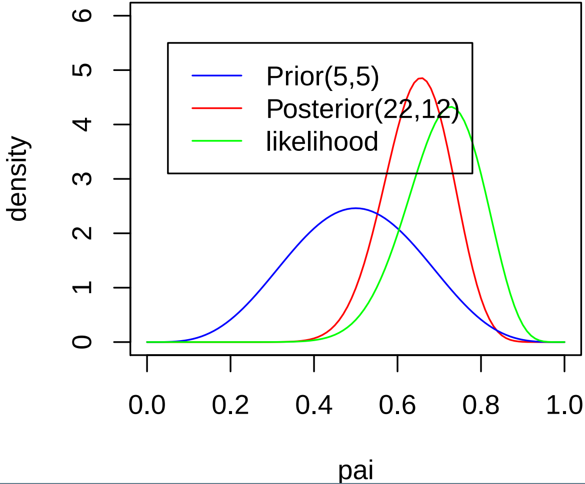
Lets continue to the previous example. You remeber 17 of 24 women say polio is contagious (so $y = 17$ and $n = 24$ where y is a realization from binomial) and you use $Beta(5, 5)$ prior; the posterior distribution is $Beta(17 + 5, 24 - 17 + 5) = Beta(22, 12)$


Prior
Posterior
density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
pai

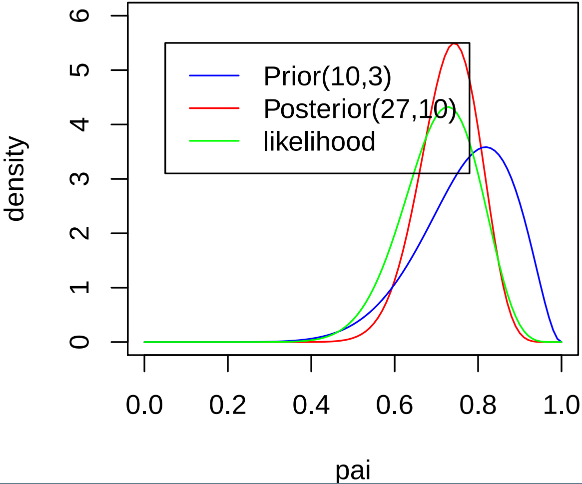
Application



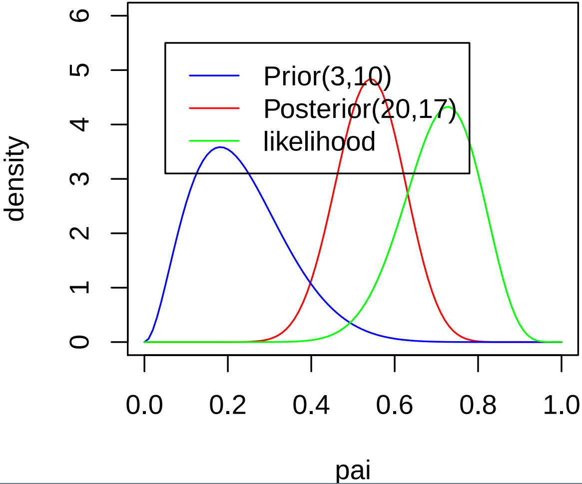
Consequence of Different Priors



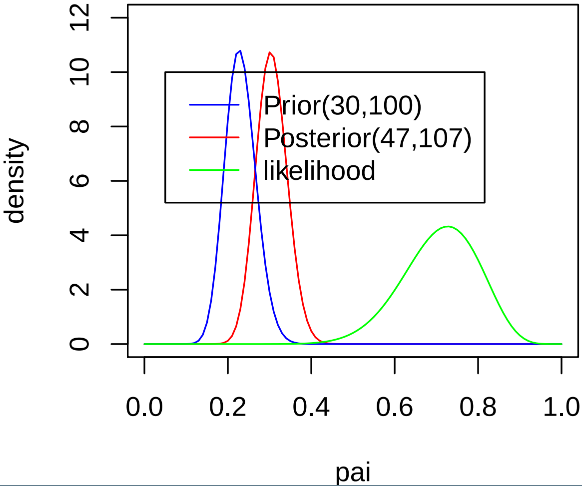
Consequence of Different Priors



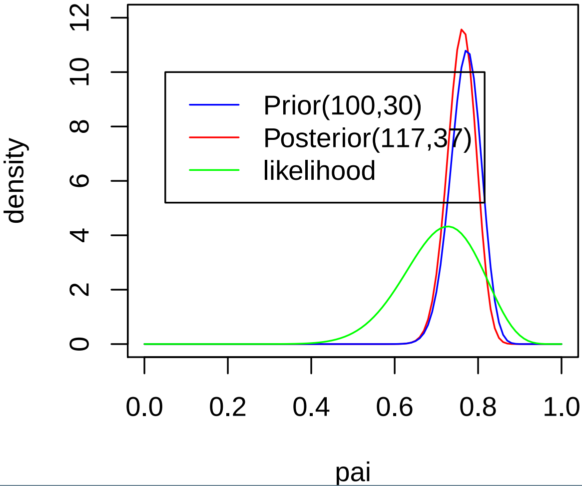
Consequence of Different Priors



Consequence of Different Priors



Consequence of Different Priors



A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference

Module: Bayesian Analysis of Normal Distribution - Part 1

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Bayesian Analysis of Mean

Bayesian estimation of the mean when variables are normally distributed with known variance.

- Benchmark: The maximum likelihood estimate for the normal distribution known variance

- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$

- The log-likelihood function:

$\log L(\mu|y) = 2n\pi - \frac{1}{2}\sum\{(y_i - \mu)^2\}$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
i.e.,

$$\frac{\partial}{\partial\mu}\log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

Benchmark

- Benchmark: The maximum likelihood estimate for the normal distribution known variance
- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$
- The log-likelihood function:

$$\log L(\mu|y) = 2n\pi - \frac{1}{2}\sum\{(y_i - \mu)^2\}$$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
i.e.,

$$\frac{\partial}{\partial\mu}\log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

- Benchmark: The maximum likelihood estimate for the normal distribution known variance

- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$

- The log-likelihood function:

$\log L(\mu|y) = 2n\pi - \frac{1}{2}\sum\{(y_i - \mu)^2\}$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
  i.e.,

$$\frac{\partial}{\partial \mu} \log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

- Benchmark: The maximum likelihood estimate for the normal distribution known variance

- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$

- The log-likelihood function:

$$\log L(\mu|y) = 2n\pi - \tfrac{1}{2}\sum\{(y_i - \mu)^2\}$$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
  i.e.,

$$\frac{\partial}{\partial\mu}\log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

- Benchmark: The maximum likelihood estimate for the normal distribution known variance

- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$

- The log-likelihood function:

$$\log L(\mu|y) = 2n\pi - \tfrac{1}{2}\sum\{(y_i - \mu)^2\}$$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
  i.e.,

$$\frac{\partial}{\partial\mu}\log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

Benchmark

- Benchmark: The maximum likelihood estimate for the normal distribution known variance

- Let $y_i \overset{iid}{\sim} N(\mu, 1) \quad , i = 1, 2, 3, ..., n$

- The log-likelihood function:

$$\log L(\mu|y) = 2n\pi - \tfrac{1}{2} \sum\{(y_i - \mu)^2\}$$

- To find MLE take the derivative of log-likelihood with respect to μ and set the equation to 0 to solve for μ.
  i.e.,

$$\frac{\partial}{\partial\mu} \log L(\mu|y) = 0$$

- MLE: $\hat{\mu} = \sum_{i=1}^{n} y_i/n$

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

Normal Model with known Variance

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

Normal Model with known Variance

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

Normal Model with known Variance

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

- Lets consider simple case when sample size $n = 1$.
- Suppose $y \sim N(\mu, \sigma_0^2)$
  - where σ_0^2 is known constant
  - μ is unknown
- What would be a good prior to choose for μ?
- What properties should the prior have?

- Why is the normal distribution a good choice?
- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?
- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?

- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?

- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?

- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?
- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

- Why is the normal distribution a good choice?
- Because the normal distribution is quite flexible...
  - The prior mean can take any value along the real line
  - The prior variance can be chosen so that our prior beliefs are (or are not) influential determinants of the posterior mean and variance.
- Because it may be a close representation of out our prior beliefs.
  - if our prior beliefs are symmetric and uni-modal about some point along the real line.
- Because it is conjugate, so the posterior will be normal.

Normal Prior

- Let $y \sim N(\mu, \sigma_0^2)$ and $\mu \sim N(m, s^2)$

$$
\begin{aligned}
p(\mu|y) &\propto f(y|\mu)p(\mu) \\
&\propto (2\pi\sigma_0^2)^{-1} \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\
&\qquad \times (2\pi s^2)^{-1} \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\}
\end{aligned}
$$

- Drop multiplicative constants

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\}
\end{aligned}
$$

Normal Prior

- Let $y \sim N(\mu, \sigma_0^2)$ and $\mu \sim N(m, s^2)$

$$\begin{aligned} p(\mu|y) &\propto f(y|\mu)p(\mu) \\ &\propto (2\pi\sigma_0^2)^{-1} \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\ &\qquad \times (2\pi s^2)^{-1} \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\} \end{aligned}$$

- Drop multiplicative constants

$$\begin{aligned} p(\mu|y) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\} \end{aligned}$$

- Let $y \sim N(\mu, \sigma_0^2)$ and $\mu \sim N(m, s^2)$

$$
\begin{aligned}
p(\mu|y) &\propto f(y|\mu)p(\mu) \\
&\propto (2\pi\sigma_0^2)^{-1} \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\
&\qquad \times (2\pi s^2)^{-1} \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\}
\end{aligned}
$$

- Drop multiplicative constants

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(y-\mu)^2\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2s^2}(\mu-m)^2\right\}
\end{aligned}
$$

Normal Prior

$$p(\mu|y) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(y^2 - 2y\mu + \mu^2) - \frac{1}{2s^2}(\mu^2 - 2m\mu + m^2)\right\}$$
$$\propto \exp\left\{ -\frac{1}{2s^2\sigma_0^2}[\sigma_0^2\mu^2 - 2\sigma_0^2\mu m + \sigma_0^2 m^2 - s^2y^2 + 2s^2y\mu + s^2\mu^2]\right\}$$

$$p(\mu|y) \propto \exp\left\{ -\frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2 + s^2) - 2\mu(\sigma_0^2 m + s^2 y) + (\sigma_0^2 m^2 - s^2y^2)]\right\}$$

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(y^2 - 2y\mu + \mu^2) - \frac{1}{2s^2}(\mu^2 - 2m\mu + m^2)\right\} \\
&\propto \exp\left\{ - \frac{1}{2s^2\sigma_0^2}[\sigma_0^2\mu^2 - 2\sigma_0^2\mu m + \sigma_0^2 m^2 \right. \\
&\qquad \left. - s^2y^2 + 2s^2y\mu + s^2\mu^2]\right\}
\end{aligned}
$$

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left\{ - \frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2 + s^2) \right. \\
&\qquad \left. - 2\mu(\sigma_0^2 m + s^2 y) + (\sigma_0^2 m^2 - s^2y^2)]\right\}
\end{aligned}
$$

$$p(\mu|y) \propto \exp\left\{ -\frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2 + s^2) - 2\mu(\sigma_0^2 m + s^2 y) + (\sigma_0^2 m^2 - s^2 y^2)] \right\}$$

Third term is free from μ. So it is part of normalizing constant. We can drop it.

$$p(\mu|y) \propto \exp\left\{ -\frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2 + s^2) - 2\mu(\sigma_0^2 m + s^2 y)] \right\}$$

$$
\begin{aligned}
p(\mu|y) \quad \propto \quad \exp\Big\{ -\frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2+s^2) \\
-2\mu(\sigma_0^2 m + s^2 y) + (\sigma_0^2 m^2 - s^2 y^2)]\Big\}
\end{aligned}
$$

Third term is free from μ. So it is part of normalizing constant. We can drop it.

$$
p(\mu|y) \quad \propto \quad \exp\Big\{ -\frac{1}{2s^2\sigma_0^2}[\mu^2(\sigma_0^2+s^2) - 2\mu(\sigma_0^2 m + s^2 y)]\Big\}
$$

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left\{-\frac{1}{2s^2\sigma_0^2}\left[\mu^2(\sigma_0^2+s^2)-2\mu(\sigma_0^2 m+s^2 y)\right]\right\} \\
&\quad \text{Multiply with } s^2\sigma_0^2 \\
&\propto \exp\left[-\frac{1}{2}\left\{\mu^2\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)-2\mu\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)\right\}\right] \\
&\quad \text{Multiply and divide by } (s^{-2}+\sigma_0^{-2})^{-1} \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left\{\frac{\mu^2\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right.\right. \\
&\qquad\qquad \left.\left.-\frac{2\mu\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right\}\right]
\end{aligned}
$$

Normal Prior

$$p(\mu|y) \propto \exp\Bigg[-\frac{1}{2}\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)\Bigg\{\frac{\mu^2\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)}{\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)} - \frac{2\mu\Big(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\Big)}{\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)}\Bigg\}\Bigg]$$

$$\propto \exp\Bigg[-\frac{1}{2}\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)\Bigg\{\mu^2 - \frac{2\mu\Big(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\Big)}{\Big(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\Big)}\Bigg\}\Bigg]$$

$$
\begin{aligned}
p(\mu|y) &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left\{\mu^2-\frac{2\mu\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left\{\mu^2-\frac{2\mu\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right\}+k-k\right]
\end{aligned}
$$

Let

$$k = \left(\frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)} \right)^2$$

$$p(\mu|y) \quad \propto \quad \exp\left[-\frac{1}{2}\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right) \left(\mu - \frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)} \right)^2 \right]$$

- Thus posterior distribution

$$p(\mu|y) \sim N(\hat{\mu}, \hat{\sigma}^2)$$

the posterior mean

$$E(\mu|y) = \hat{\mu} = \frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)}$$

- Thus posterior distribution

$$p(\mu|y) \sim N(\hat{\mu}, \hat{\sigma}^2)$$

the posterior mean

$$E(\mu|y) = \hat{\mu} = \frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)}$$

the posterior variance

$$Var(\mu|y) = \hat{\sigma}^2 = \left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

- The posterior mean is the weighted average of the data and prior mean, i.e.,

$$E(\mu|y) = \hat{\mu} = \frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)}$$
$$= \omega\ m + (1-\omega)y$$

where

$$\omega = \frac{\left(\frac{1}{s^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)} = \frac{\sigma_0^2}{\sigma_0^2 + s^2}$$

- The posterior mean is the weighted average of the data and prior mean, i.e.,

$$
\begin{aligned}
E(\mu|y) = \hat{\mu} &= \frac{\left(\frac{m}{s^2} + \frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)} \\
&= \omega\ m + (1-\omega)y
\end{aligned}
$$

where

$$
\omega = \frac{\left(\frac{1}{s^2}\right)}{\left(\frac{1}{s^2} + \frac{1}{\sigma_0^2}\right)} = \frac{\sigma_0^2}{\sigma_0^2 + s^2}
$$

With n observation

If we have more than one observation of the random variable y from the normal distribution with unknown mean and known variance, it is a trivial matter to extend this result to show that:

$$p(\mu|y_1, ..., y_n) \sim N(\hat{\mu}, \hat{\sigma}^2)$$

where

$$\hat{\mu} = E(\mu|y_1, ..., y_n) = \frac{\left(\frac{m}{s^2} + \frac{n\bar{y}}{\sigma_0^2}\right)}{\left(\frac{1}{s^2} + \frac{n}{\sigma_0^2}\right)}$$

and

$$\hat{\sigma}^2 = Var(\mu|y_1, ..., y_n) = \left(\frac{1}{s^2} + \frac{n}{\sigma_0^2}\right)^{-1}$$

- The posterior mean:

$$E(\mu|y_1, y_2, ..., y_n) = \omega m + (1-\omega)\bar{y}$$

where $\omega = \dfrac{\left(\frac{1}{s^2}\right)}{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)} = \frac{\sigma_0^2}{\sigma_0^2+ns^2}$ and $1-\omega = \frac{ns^2}{\sigma_0^2+ns^2}$

Note

Clearly, as $n \to \infty$ then $\omega \to 0$, therefore posterior mean tends towards mle, i.e., effect of prior washes away as sample size become larger.

- The posterior mean:

$$E(\mu|y_1, y_2, ..., y_n) = \omega m + (1-\omega)\bar{y}$$

where $\omega = \dfrac{\left(\frac{1}{s^2}\right)}{\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)} = \frac{\sigma_0^2}{\sigma_0^2+ns^2}$ and $1-\omega = \frac{ns^2}{\sigma_0^2+ns^2}$

Note

Clearly, as $n \to \infty$ then $\omega \to 0$, therefore posterior mean tends towards mle, i.e., effect of prior washes away as sample size become larger.

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Based on national studies, we "know" that the standard deviation of y is 13. We will use this information in our favour.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, 13)$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Based on national studies, we "know" that the standard deviation of y is 13. We will use this information in our favour.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, 13)$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Based on national studies, we "know" that the standard deviation of y is 13. We will use this information in our favour.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, 13)$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Based on national studies, we "know" that the standard deviation of y is 13. We will use this information in our favour.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, 13)$

Application

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m, s^2)$
- Why might it be reasonable to choose a normal prior for μ? Why might it be useful?
- What are the problems with using a normal random variable to describe the blood pressure data?
- What values should we choose from m and s^2?
- What is the posterior distribution of μ?

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m, s^2)$
- Why might it be reasonable to choose a normal prior for μ? Why might it be useful?
- What are the problems with using a normal random variable to describe the blood pressure data?
- What values should we choose from m and s^2?
- What is the posterior distribution of μ?

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m, s^2)$
- Why might it be reasonable to choose a normal prior for μ? Why might it be useful?
- What are the problems with using a normal random variable to describe the blood pressure data?
- What values should we choose from m and s^2?
- What is the posterior distribution of μ?

Application

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m, s^2)$
- Why might it be reasonable to choose a normal prior for μ? Why might it be useful?
- What are the problems with using a normal random variable to describe the blood pressure data?
- What values should we choose from m and s^2?
- What is the posterior distribution of μ?

Application

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m, s^2)$
- Why might it be reasonable to choose a normal prior for μ? Why might it be useful?
- What are the problems with using a normal random variable to describe the blood pressure data?
- What values should we choose from m and s^2?
- What is the posterior distribution of μ?

- $y_i \sim N(\mu, 13^2)$ and $\mu \sim N(m = 0, s^2 = 1000)$, $\bar{y} = 128$ and $n = 20$

- What is the posterior distribution of μ?

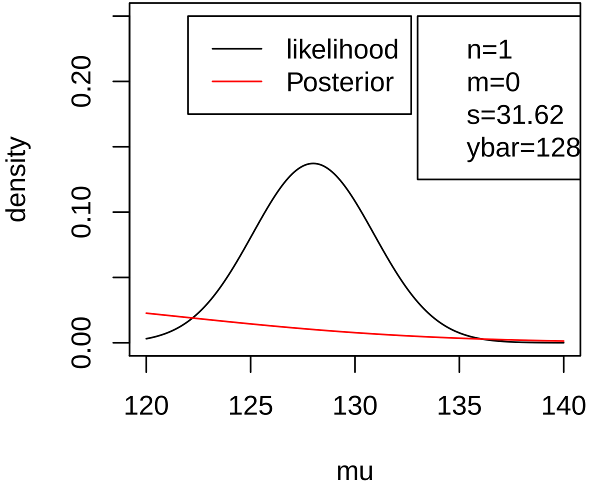
$$p(\mu|y_1, ..., y_n) \sim N(\hat{\mu}, \hat{\sigma}^2)$$

where

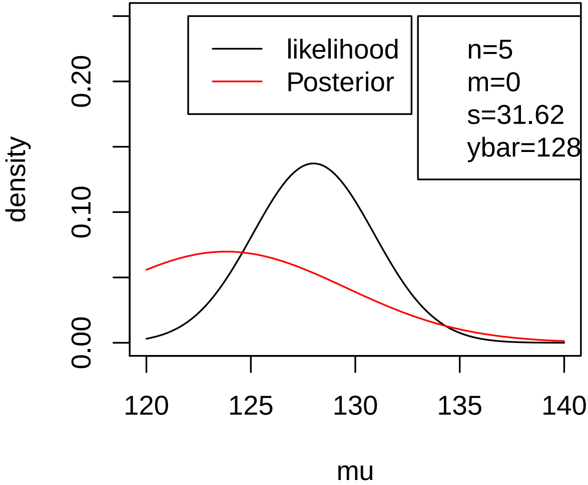
$$\hat{\mu} = E(\mu|y_1, ..., y_n) = \frac{\left(\frac{0}{1000} + \frac{20(128)}{13^2}\right)}{\left(\frac{1}{1000} + \frac{20}{13^2}\right)} \approx 126.93$$

and

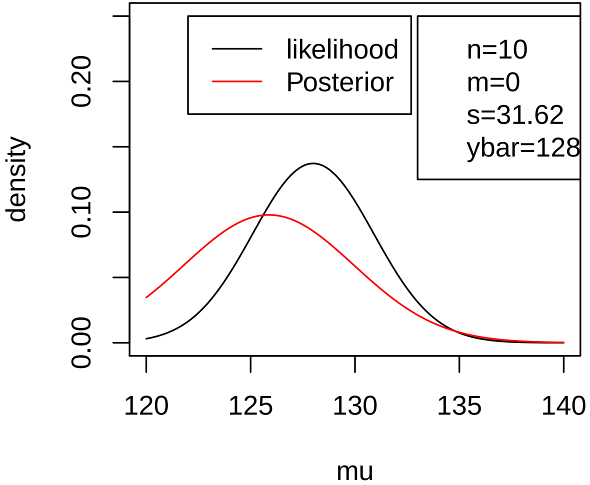
$$\hat{\sigma}^2 = Var(\mu|y_1, ..., y_n) = \left(\frac{1}{1000} + \frac{20}{13^2}\right)^{-1} \approx 8.38$$


likelihood
Posterior
n=1
m=0
s=31.62
ybar=128
density
0.00
0.10
0.20
120
125
130
135
140
mu

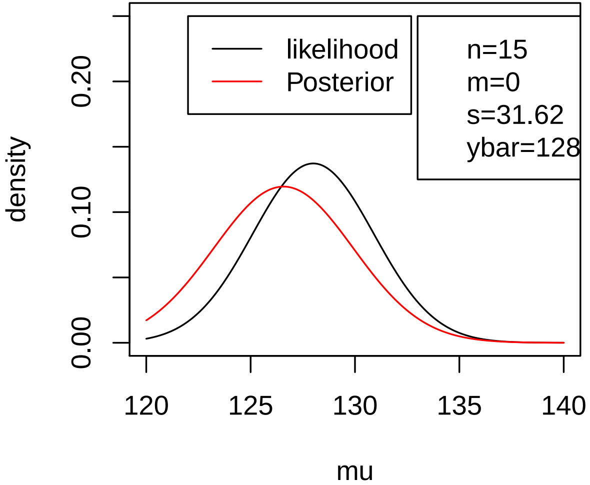
Application



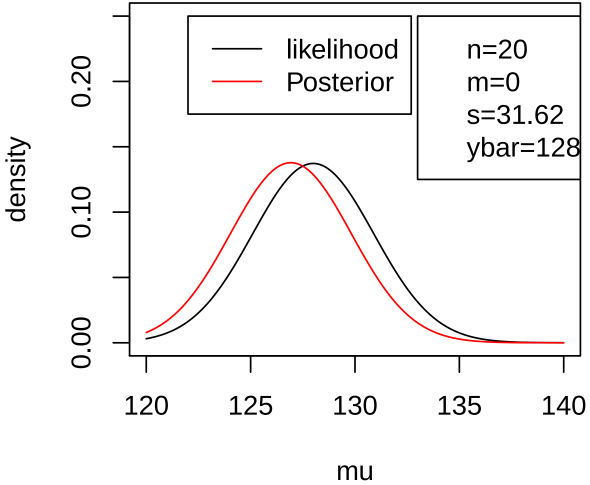
Application



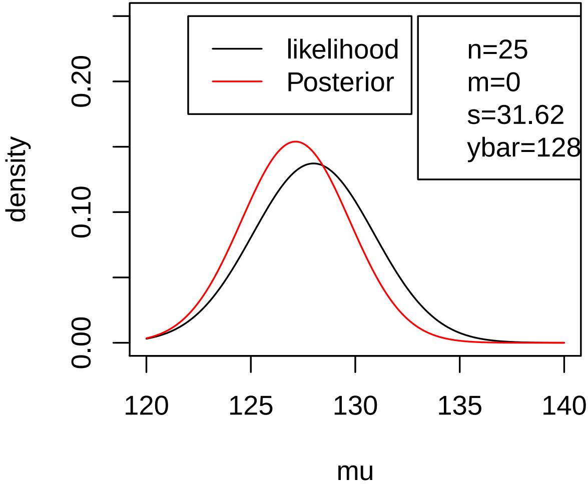
Application

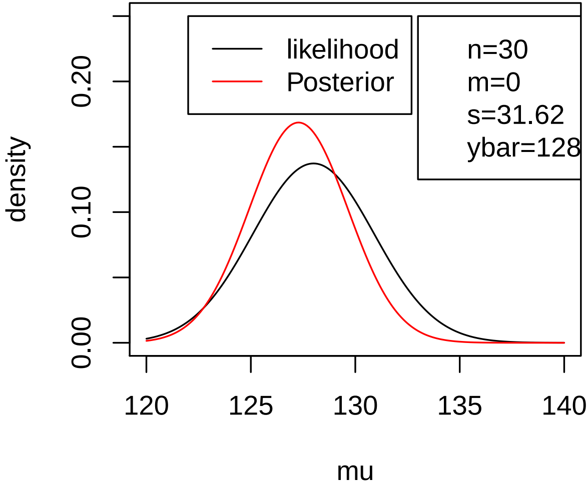


Application



Application




likelihood
Posterior
n=30
m=0
s=31.62
ybar=128
density
0.00
0.10
0.20
120
125
130
135
140
mu

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference

Module: Bayesian Analysis of Normal Distribution - Part 2

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Bayesian Analysis of Mean

Bayesian estimation of the mean when variables are normally distributed and variance is also unknown.

- In this module, we will consider Non-informative Improper prior for unknown mean and variance

Normal Model with unknown Variance

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.

- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.

- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

Normal Model with unknown Variance

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.
- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.
- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.

- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.

- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

Normal Model with unknown Variance

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.

- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.

- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

Normal Model with unknown Variance

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.

- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.

- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

Normal Model with unknown Variance

- Suppose $y_i \sim N(\mu, \sigma^2)$
  - where σ^2 and μ are unknown random variable
- This is the first example of multi-parameter model.

- The Bayesian setup still look familiar:

$$p(\mu, \sigma | y_1, ..., y_n) \propto p(y|\mu, \sigma)p(\mu, \sigma)$$

- we would like to make inferences about the marginal distributions $p(\mu|y)$ and $p(|y)$, where $y = (y_1, ..., y_n)$; rather than the joint distribution $p(\mu, \sigma|y)$.

- Ultimately we would like to find:

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

- Note that the equation

$$p(\mu|y) = \int p(\mu, \sigma|y) d\sigma$$

cen be presented as:

$$p(\mu|y) = \int p(\mu|\sigma, y) p(\sigma|y) d\sigma$$

Different Approach to Choose Priors

Classical Bayesians

The prior is a necessary evil. Choose priors that interject the least information possible.

Modern Parametric Bayesians

The prior is a useful convenience. Choose prior distributions with desirable properties (e.g. conjugacy). Given a distributional choice, prior parameters are chosen to interject the least information possible.

Subjective Bayesians

The prior is a summary of old beliefs Choose prior distributions based on previous knowledge - either the results of earlier studies or non-scientific opinion.

Different Approach to Choose Priors

Classical Bayesians

The prior is a necessary evil. Choose priors that interject the least information possible.

Modern Parametric Bayesians

The prior is a useful convenience. Choose prior distributions with desirable properties (e.g. conjugacy). Given a distributional choice, prior parameters are chosen to interject the least information possible.

Subjective Bayesians

The prior is a summary of old beliefs Choose prior distributions based on previous knowledge - either the results of earlier studies or non-scientific opinion.

Different Approach to Choose Priors

Classical Bayesians

The prior is a necessary evil. Choose priors that interject the least information possible.

Modern Parametric Bayesians

The prior is a useful convenience. Choose prior distributions with desirable properties (e.g. conjugacy). Given a distributional choice, prior parameters are chosen to interject the least information possible.

Subjective Bayesians

The prior is a summary of old beliefs Choose prior distributions based on previous knowledge - either the results of earlier studies or non-scientific opinion.

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables.
- What prior distribution would you choose to represent the absence of any knowledge in this instance?
- What if we assumed that the two parameters were independent, so $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables.

- What prior distribution would you choose to represent the absence of any knowledge in this instance?

- What if we assumed that the two parameters were independent, so $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables.

- What prior distribution would you choose to represent the absence of any knowledge in this instance?

- What if we assumed that the two parameters were independent, so $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables. What prior should we choose?

- if $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ - one option would be to assume uniform prior distributions for both parameters. Thus

$$\begin{aligned} p(\mu) &\propto c \quad \text{for} \quad -\infty < \mu < \infty \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{for} \quad 0 < \sigma^2 < \infty \end{aligned}$$

- And the joint density would be: $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$?

- Are these distributions proper?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables. What prior should we choose?

- if $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ - one option would be to assume uniform prior distributions for both parameters. Thus

$$\begin{aligned} p(\mu) &\propto c \quad \text{for} \quad -\infty < \mu < \infty \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{for} \quad 0 < \sigma^2 < \infty \end{aligned}$$

- And the joint density would be: $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$?

- Are these distributions proper?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables. What prior should we choose?

- if $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ - one option would be to assume uniform prior distributions for both parameters. Thus

$$\begin{aligned} p(\mu) &\propto c \quad \text{for} \quad -\infty < \mu < \infty \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{for} \quad 0 < \sigma^2 < \infty \end{aligned}$$

- And the joint density would be: $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$?

- Are these distributions proper?

The Classical Bayesian: normal model with unknown mean and variance

- The $y \sim N(\mu, \sigma^2)$ where μ and σ are both unknown and random variables. What prior should we choose?

- if $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ - one option would be to assume uniform prior distributions for both parameters. Thus

$$\begin{aligned} p(\mu) &\propto c \quad \text{for} \quad -\infty < \mu < \infty \\ p(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{for} \quad 0 < \sigma^2 < \infty \end{aligned}$$

- And the joint density would be: $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$?

- Are these distributions proper?

The Classical Bayesian: normal model with unknown mean and variance

- Let $y_i \sim N(\mu, \sigma^2)$ where $i = 1, 2, ..., n$ and $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$

- The posterior distribution is:

$$
\begin{aligned}
p(\mu, \sigma^2|y) &\propto p(y|\mu, \sigma^2) \times p(\mu, \sigma^2) \\
&\propto \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y} - \mu)^2\right] \\
&\quad \times \frac{1}{\sigma^2}
\end{aligned}
$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $y = (y_1, ..., y_n)$

- It can be shown that the conditional posterior distribution

$$p(\mu|\sigma^2, y) \sim N(\bar{y}, \sigma^2/n)$$

The Classical Bayesian: normal model with unknown mean and variance

- Let $y_i \sim N(\mu, \sigma^2)$ where $i = 1, 2, ..., n$ and $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$

- The posterior distribution is:

$$
\begin{aligned}
p(\mu, \sigma^2 | y) &\propto p(y|\mu, \sigma^2) \times p(\mu, \sigma^2) \\
&\propto \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y} - \mu)^2\right] \\
&\qquad \times \frac{1}{\sigma^2}
\end{aligned}
$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $y = (y_1, ..., y_n)$

- It can be shown that the conditional posterior distribution

$$p(\mu|\sigma^2, y) \sim N(\bar{y}, \sigma^2/n)$$

The Classical Bayesian: normal model with unknown mean and variance

- Let $y_i \sim N(\mu, \sigma^2)$ where $i = 1, 2, ..., n$ and $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$

- The posterior distribution is:

$$\begin{aligned} p(\mu, \sigma^2 | y) &\propto p(y|\mu, \sigma^2) \times p(\mu, \sigma^2) \\ &\propto \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y} - \mu)^2\right] \\ &\qquad \times \frac{1}{\sigma^2} \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $y = (y_1, ..., y_n)$

- It can be shown that the conditional posterior distribution

$$p(\mu|\sigma^2, y) \sim N(\bar{y}, \sigma^2/n)$$

The Classical Bayesian: normal model with unknown mean and variance

- The marginal posterior distribution of μ is

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2$$

- It can be shown that

$$p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$$

- Or more conviniently,

$$p\left(\frac{\sqrt{n}(\mu - \bar{y})}{s}\right) \sim t_{n-1}$$

The Classical Bayesian: normal model with unknown mean and variance

- The marginal posterior distribution of μ is

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2$$

- It can be shown that

$$p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$$

- Or more conviniently,

$$p\left(\frac{\sqrt{n}(\mu - \bar{y})}{s}\right) \sim t_{n-1}$$

The Classical Bayesian: normal model with unknown mean and variance

- The marginal posterior distribution of μ is

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2$$

- It can be shown that

$$p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$$

- Or more conviniently,

$$p\left(\frac{\sqrt{n}(\mu - \bar{y})}{s}\right) \sim t_{n-1}$$

- The marginal posterior distribution of σ^2 is

$$p(\sigma^2|y) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y}-\mu)^2\right] d\mu$$

- It can be shown that it follows scaled-inverse χ^2 distribution

$$p(\sigma^2|y) \sim \text{Inv-}\chi^2(n-1, s^2)$$

- Or more conviniently,

$$p(\sigma^2|y) \sim \text{Inv-Gamma}((n-1)/2, (n-1)s^2/2)$$

The Classical Bayesian: normal model with unknown mean and variance

- The marginal posterior distribution of σ^2 is

$$p(\sigma^2|y) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y}-\mu)^2\right] d\mu$$

- It can be shown that it follows scaled-inverse χ^2 distribution

$$p(\sigma^2|y) \sim \text{Inv-}\chi^2(n-1, s^2)$$

- Or more conviniently,

$$p(\sigma^2|y) \sim \text{Inv-Gamma}((n-1)/2, (n-1)s^2/2)$$

The Classical Bayesian: normal model with unknown mean and variance

- The marginal posterior distribution of σ^2 is

$$p(\sigma^2|y) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2 + n(\bar{y}-\mu)^2\right] d\mu$$

- It can be shown that it follows scaled-inverse χ^2 distribution

$$p(\sigma^2|y) \sim \text{Inv-}\chi^2(n-1, s^2)$$

- Or more conviniently,

$$p(\sigma^2|y) \sim \text{Inv-Gamma}((n-1)/2, (n-1)s^2/2)$$

- Big point to note here:

- Though prior is flat but improper - posterior is a proper probability distribution.

- Hence proper statistical inference can be drawn from it.

Two Different methods to sample from the posterior in this case

Method 1

Sample directly from each of the two marginal distributions.

Method 2

Two stages:

1. Sample a value from the marginal distribution of $\sigma^2|y$

2. Sample a value from the marginal distribution of $\mu|\sigma^2, y$

Two Different methods to sample from the posterior in this case

Method 1

Sample directly from each of the two marginal distributions.

Method 2

Two stages:

1. Sample a value from the marginal distribution of $\sigma^2|y$
2. Sample a value from the marginal distribution of $\mu|\sigma^2, y$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, \sigma^2)$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, \sigma^2)$

Application

What is the average blood pressure reading of a sub-population of 20 adult men?

- Let y_i denote the blood pressure of individual i and assume that y_i is normally distributed.
- Let μ denote the average blood pressure of the group.
- Thus $y_i \sim N(\mu, \sigma^2)$

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution is $p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$
- By the properties of t-distribution which we can find as $E(\mu|y) = \bar{y} = 128$ and $Var(\mu|y) = \left(\frac{s^2}{n-2}\right) = 3.27$

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution is $p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$
- By the properties of t-distribution which we can find as $E(\mu|y) = \bar{y} = 128$ and $Var(\mu|y) = \left(\frac{s^2}{n-2}\right) = 3.27$

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution is $p(\mu|y) \sim t_{n-1}(\bar{y}, s^2/n)$
- By the properties of t-distribution which we can find as $E(\mu|y) = \bar{y} = 128$ and $Var(\mu|y) = \left(\frac{s^2}{n-2}\right) = 3.27$

- Analytical Results:

$$E(\mu|y) = \bar{y} = 128 \text{ and } Var(\mu|y) = \left(\frac{s^2}{n-2}\right) = 3.27$$

- Numerical Results:

```
Summary of mu :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  121.3   126.8   128.0   128.0   129.2   135.9
Variance of mu :
[1] 3.285445
Sd of mu :
[1] 1.81258
95% CI of mu :
    2.5%    97.5%
124.3952 131.5632
```

- Analytical Results:

$$E(\mu|y) = \bar{y} = 128 \text{ and } Var(\mu|y) = \left(\frac{s^2}{n-2}\right) = 3.27$$

- Numerical Results:

```
Summary of mu :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  121.3   126.8   128.0   128.0   129.2   135.9
Variance of mu :
[1] 3.285445
Sd of mu :
[1] 1.81258
95% CI of mu :
    2.5%    97.5%
124.3952 131.5632
```

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution of σ^2 is $p(\sigma^2|y) \sim$ Inv-$\chi^2(n-1, s^2)$
- By the properties of scales inv χ^2-distribution we can find as $E(\sigma^2|y) = s^2 \left(\frac{n-1}{n-3}\right) = 64.9$

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution of σ^2 is $p(\sigma^2|y) \sim$ Inv-$\chi^2(n-1, s^2)$
- By the properties of scales inv χ^2-distribution we can find as $E(\sigma^2|y) = s^2\left(\frac{n-1}{n-3}\right) = 64.9$

- The sample average is $\bar{y} = 128$ and sample sd is $s = 7.67$
- The marginal distribution of σ^2 is $p(\sigma^2|y) \sim \text{Inv-}\chi^2(n-1, s^2)$
- By the properties of scales inv χ^2-distribution we can find as $E(\sigma^2|y) = s^2\left(\frac{n-1}{n-3}\right) = 64.9$

Application

- Analytical Result:

$$E(\sigma^2|y) = s^2\left(\frac{n-1}{n-3}\right) = 64.9$$

- Numerical Result:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.09	49.04	60.49	65.36	76.35	362.80

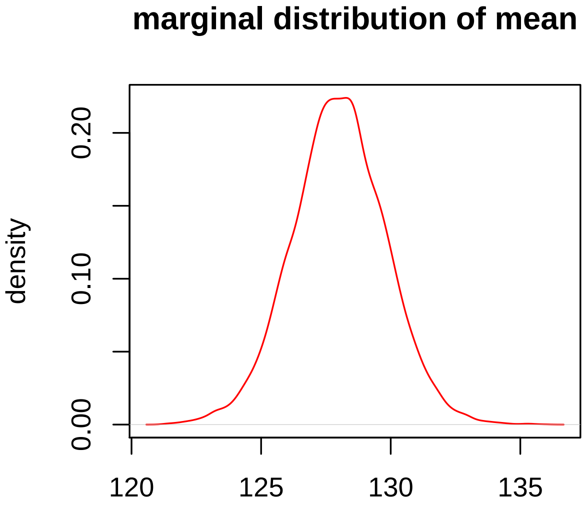
- Analytical Result:

$$E(\sigma^2|y) = s^2\left(\frac{n-1}{n-3}\right) = 64.9$$

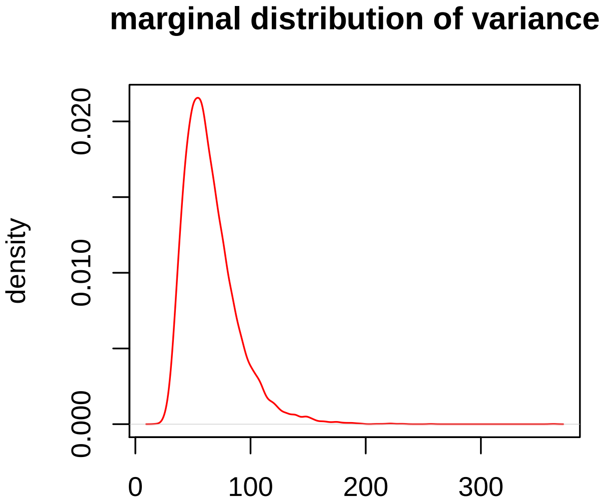
- Numerical Result:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.09	49.04	60.49	65.36	76.35	362.80

plot of posterior densities from method 1



plot of posterior densities from method 1



Numerical Method 2

Two stages:

1. Sample a value from the marginal distribution of $\sigma^2|y$
2. Sample a value from the marginal distribution of $\mu|\sigma^2, y$

Comparing Analytical Results with Numerical Method 2

Numerical Method 2

Two stages:

1. Sample a value from the marginal distribution of $\sigma^2|y$
2. Sample a value from the marginal distribution of $\mu|\sigma^2, y$

Comparing Analytical Results with Numerical Method 2

```
Summary of mu :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  118.7   126.8   128.0   128.0   129.2   136.4
Variance of mu :
[1] 3.147524
Sd of mu :
[1] 1.774126
95% of mu :
    2.5%    97.5%
124.5105 131.5384
```

Comparision between Analytical and Numerical Results

	Analytical	Numerical Method I	Numerical Method II
$E(\mu\|y)$	128	128.0	128.0
$V(\mu\|y)$	3.27	3.29	3.15
$E(\sigma^2\|y)$	64.9	65.4	65.4

Bayesian Conjugate Priors for the Normal Distribution

1 Conjugate Prior

Definition:

Let X be a random variable having density $f(x|\theta)$ indexed by θ.
Let $\mathcal{F}$=$\{f(x|\theta)$; x $\in \mathcal{X}$, $\theta \in \Theta\}$ be a class of density functions.
A class $\mathcal{P}$=$\{\pi(\theta)$; $\theta \in \Theta\}$ of prior distributions is said to be a conjugate family for $\mathcal{F}$ if the posterior distribution $\pi(\theta|x)$ is in the class $\mathcal{P}$ for all x $\in \mathcal{X}$ and for all $\pi \in \mathcal{P}$

Example: Suppose X $\sim N(\mu, \sigma^2)$, $\mu \in \Theta \equiv \Re$

$$f(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \Re$$

Suppose $\mathcal{P}$=$\{ N(\alpha, \beta^2), \quad \alpha \in \Re ; \beta > 0\}$, i.e

$$\pi(\mu) = \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{1}{2\beta^2}(\mu-\alpha)^2}, \quad \mu \in \Re$$

Then $\pi(\mu|x) \sim N(\frac{\beta^2 x}{\beta^2+\sigma^2} + \frac{\sigma^2\alpha}{\beta^2+\sigma^2}, \frac{1}{\frac{1}{\sigma^2}+\frac{1}{\beta^2}})$
i.e $\pi(\mu|x) \in \mathcal{P}$

Hence $\mathcal{P}$ is conjugate for $\mathcal{F}$,

where $\mathcal{F}$=$\{N(\mu,\sigma^2)$, $\mu \in \Re$, $\sigma^2 > 0$ is known$\}$

Note:- If $X_1, X_2, ...X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \Re$ and σ^2 is known

Then $\mathcal{F}$=$\{f(x|\mu) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}, \quad \mu \in \Re\}$

If $\mathcal{P}$=$\{ N(\alpha, \beta^2), \quad \alpha \in \Re ; \beta > 0\}$,

Then $\pi(\mu|\mathbf{x}) \sim N(\frac{\beta^2\bar{x} + \frac{\sigma^2}{n}\alpha}{\beta^2+\frac{\sigma^2}{n}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\beta^2}})$

Let $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, both μ and σ^2 are unknown

Let us consider the following prior distributions

$\pi(\mu|\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}} \; e^{-\frac{1}{2\frac{\sigma^2}{s_0}}(\mu-\delta)^2}$ $\longrightarrow$ Normal Prior

$\pi(\sigma^2) \propto (\sigma^2)^{(\alpha+1)} \; e^{-\frac{\beta}{\sigma^2}}$ $\longrightarrow$ Inverse Gamma Prior

<u>To find the Posterior Distribution</u>

Posterior distribution of (μ, σ^2)

$\pi(\mu, \sigma^2|\mathbf{x}) \propto L(\mu, \sigma^2|\mathbf{x}) \;\; \pi(\mu|\sigma^2)\pi(\sigma^2)$

$$\propto (\frac{1}{\sigma\sqrt{2\pi}})^n \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \; \frac{1}{\sigma} \; e^{-\frac{1}{2\frac{\sigma^2}{s_0}}(\mu-\delta)^2} \; \frac{1}{\sigma^2} \; e^{-\frac{\beta}{\sigma^2}}$$

<u>To find the Posterior Distribution of σ^2</u>

Coeff of μ^2= $-\frac{n}{2\sigma^2} - \frac{s_0}{2\sigma^2} = -(\frac{n+s_0}{2\sigma^2})$

Coeff of $\mu = \frac{n\bar{x}}{\sigma^2} + \frac{s_0\delta}{\sigma^2} = (\frac{n\bar{x}+s_0\delta}{\sigma^2})$

Now,

$$\pi(\mu, \sigma^2|\mathbf{x}) = const\,(\frac{1}{\sigma^2})^{\frac{n}{2}+\alpha+1+\frac{1}{2}} \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} {x_i}^2 - \frac{1}{2\frac{\sigma^2}{s_0}}\delta^2 - \frac{\beta}{\sigma^2}} \quad e^{-(\frac{n+s_0}{2\sigma^2})\mu^2 + (\frac{n\bar{x}+s_0\delta}{\sigma^2})\mu}$$

As

$$-\frac{n+s_0}{2\sigma^2}[\mu^2 - 2\,\frac{n\bar{x}+s_0\delta}{n+s_0}\,\mu]$$
$$= -\frac{n+s_0}{2\sigma^2}(\mu - \frac{n\bar{x}+s_0\delta}{n+s_0})^2 \quad + \quad \frac{(n\bar{x}+s_0\delta)^2}{2\sigma^2(n+s_0)}$$

We have

$$\pi(\mu, \sigma^2|\mathbf{x}) = const\,(\frac{1}{\sigma^2})^{\frac{n}{2}+\alpha+1+\frac{1}{2}} \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} {x_i}^2 - \frac{1}{2\frac{\sigma^2}{s_0}}\delta^2 - \frac{\beta}{\sigma^2} + \frac{(n\bar{x}+s_0\delta)^2}{2\sigma^2(n+s_0)}} \quad e^{-\frac{1}{2\frac{\sigma^2}{n+s_0}}(\mu - \frac{n\bar{x}+s_0\delta}{n+s_0})^2}$$

$$\pi(\sigma^2|\mathbf{x}) = \int_{-\infty}^{\infty} \pi(\mu, \sigma^2|\mathbf{x})\, d\mu$$

$$= const\,(\tfrac{1}{\sigma^2})^{\frac{n}{2}+\alpha+1+\frac{1}{2}}\; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} {x_i}^2 - \frac{1}{2\frac{\sigma^2}{s_0}}\delta^2 - \frac{\beta}{\sigma^2} + \frac{(n\bar{x}+s_0\delta)^2}{2\sigma^2(n+s_0)}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{1}{2\frac{\sigma^2}{n+s_0}}(\mu-\frac{n\bar{x}+s_0\delta}{n+s_0})^2}\, d\mu}_{\sqrt{2\pi}\frac{\sigma}{\sqrt{n+s_0}}}$$

$$\longrightarrow Inverse\ Gamma\,(\tfrac{n}{2}+\alpha, \tfrac{1}{2}\sum_{i=1}^{n} {x_i}^2 - \tfrac{1}{2}\tfrac{\delta^2}{s_0} - \beta + \tfrac{(n\bar{x}+s_0\delta)^2}{2(n+s_0)})$$

Posterior Distribution of μ given σ^2

$\pi(\mu|\sigma^2, \mathbf{x}) = \frac{\pi(\mu,\sigma^2|\mathbf{X})}{\pi(\sigma^2|\mathbf{X})}$

Thus we have

$\pi(\mu|\sigma^2, \mathbf{x}) = \frac{1}{\sqrt{2\pi}\left(\frac{\sigma}{\sqrt{n+s_0}}\right)}\, e^{-\frac{1}{2\left(\frac{\sigma^2}{n+s_0}\right)}\{\mu-\frac{n\bar{x}+s_0\delta}{n+s_0}\}^2}$

i.e

$\pi(\mu|\sigma^2, \mathbf{x}) \sim N(\frac{n\bar{x}+s_0\delta}{n+s_0}, \frac{\sigma^2}{n+s_0})$

$\longrightarrow Normal\, distribution$

Hence the class of prior distributions $\{\pi(\mu|\sigma^2),\, \mu \in \Re\}$ *and* $\{\pi(\sigma^2),\, \sigma^2 > 0\}$ as considered are conjugate.

Also,

$$\pi(\mu|\mathbf{x}) = \int_0^\infty \pi(\mu, \sigma^2|\mathbf{x})\, d\sigma^2$$

$$= \int_0^\infty \underbrace{\pi(\mu|\sigma^2, \mathbf{x})}_{\text{Normal}} \quad \underbrace{\pi(\sigma^2|\mathbf{x})}_{\text{Inverse Gamma}} \quad d\sigma^2$$

$\longrightarrow$leads to a t distribution with df=$(2\alpha + 1 + n)$ multiplied by some scale parameter

2 Posterior Predictive Distribution

If $X_1, X_2, ..., X_n \stackrel{iid}{\sim} f(x|\theta)$
$\pi(\theta) \longrightarrow$prior distribution of θ
$\pi(\theta|\mathbf{x}) \longrightarrow$posterior distribution of θ given the data $\mathbf{x}$

The posterior predictive distribution of a new observation x^* is given by

$p(x^*|\mathbf{x}) = \int_{\theta\in\Theta} f(x^*|\theta)\,\pi(\theta|\mathbf{x})\,d\theta$

Example:

Suppose $X_1, X_2, ..., X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ both μ and σ^2 are unknown
$\pi(\mu|\sigma^2) = N(\delta\,,\,\frac{\sigma^2}{s_0})$
$\pi(\sigma^2) = IG(\alpha\,,\,\beta)$

Then, $\pi(\mu, \sigma^2|\mathbf{x})$= $\pi(\mu|\sigma^2,\,\mathbf{x})\;\pi(\sigma^2|\mathbf{x})$

$$= N(\tfrac{n\bar{x}+s_0\delta}{n+s_0}, \tfrac{\sigma^2}{n+s_0})\;IG\,(\alpha + \tfrac{n}{2}, \beta^*)\text{ , where}$$

β^*= $\frac{1}{2}\sum_{i=1}^{n} x_i{}^2 - \frac{1}{2}\frac{\delta^2}{s_0} - \beta + \frac{(n\bar{x}+s_0\delta)^2}{2(n+s_0)}$

Let $\gamma = \frac{(n\bar{x}+s_0\delta)^2}{2(n+s_0)}$ and $b^2 = \frac{\sigma^2}{n+s_0}$

The posterior predictive distribution of a new observation x^* is

$$p(x^*|\mathbf{x}) = \int_\mu \int_{\sigma^2} \underbrace{f(x^*|\mu,\sigma^2)}_{\text{N}(\mu,\sigma^2)}\;\underbrace{\pi(\mu|\sigma^2,\mathbf{x})}_{\text{N}(\gamma,b^2)}\;\underbrace{\pi(\sigma^2|\mathbf{x})}_{\text{Inverse Gamma}}\;d\mu\,d\sigma^2$$

$\longrightarrow$ This leads to a t-distribution with df= $n + 2(\alpha + 1)$ multiplied by some scale parameter

2.1 Drawing samples from the Posterior Predictive Distribution

Method 1 : -

We draw samples from a t-distribution with df= $n+ 2(\alpha+1)$ multiplied by some scale parameter (which is a function of the hyper parameters), by evaluating the exact analytic form of the distribution

Alternative Method : -

Now, $p(x^*|\mathbf{x}) = \int_\mu \int_{\sigma^2} N(\mu, \sigma^2)\ N(\gamma, b^2)\ IG\,(\alpha + \frac{n}{2},\, \beta^*)\, d\sigma^2\ d\mu$

where $\gamma\,, b^2\,, \beta^*$ are as before

Method

1⟩ Draw ${\sigma_j}^2$ from $\pi(\sigma^2|\mathbf{x}) = \mathrm{IG}\,(\alpha + \frac{n}{2},\, \beta^*)$
2⟩ Draw μ_j from $\pi(\mu|{\sigma_j}^2\,, \mathbf{x}) = \mathrm{N}\,(\gamma,\, \frac{{\sigma_j}^2}{n+s_0})$
3⟩ Draw ${x_i}^*$ from $\mathrm{N}(\mu_j, {\sigma_j}^2)$

$\longrightarrow$ Repeat this to get desired number of samples from the required t-distribution

2.2 Highest Posterior Density Credible Set

X is a random variable with density $f(x|\theta)$ indexed by θ
$\pi(\theta) \longrightarrow$ prior distribution of θ
$\pi(\theta|\mathbf{x}) \longrightarrow$ posterior distribution of θ given the data $\mathbf{x}$
The 100(1-α)% HPD credible set for θ is the subset C of Θ, where

C=$\{\theta \in \Theta : \pi(\theta|\mathbf{x}) \geq \mathrm{k}(\alpha)\}$, where, k($\alpha$) is such that

$\int_{k(\alpha)}^{\infty} \pi(\theta|\mathbf{x})\, d\theta \geq 1 - \alpha$

Example:

Suppose $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ both μ and σ^2 are unknown

We want to find the 100 (1-α)% HPD credible set for μ

As usual we have,

$\pi(\mu|\sigma^2) = N(\delta\,,\,\frac{\sigma^2}{s_0})$

$\pi(\sigma^2) = IG(\alpha\,,\,\beta)$

$\pi(\mu, \sigma^2|\mathbf{x})= \pi(\mu|\sigma^2,\, \mathbf{x})\; \pi(\sigma^2|\mathbf{x})$

$\quad = N(\gamma, b^2)\; IG\,(\alpha + \frac{n}{2}, \beta^*)$, where

$\beta^*= \frac{1}{2}\sum_{i=1}^{n} {x_i}^2 - \frac{1}{2}\frac{\delta^2}{s_0} - \beta + \frac{(n\bar{x}+s_0\delta)^2}{2(n+s_0)}$

$\gamma = \frac{(n\bar{x}+s_0\delta)^2}{2(n+s_0)}$ and $b^2 = \frac{\sigma^2}{n+s_0}$

$\pi(\mu|\mathbf{x}) = \int_0^\infty \pi(\mu,\, \sigma^2|\,\mathbf{x})\, d\sigma^2$

$\longrightarrow$ t-distribution with df=2α+n+1

Suppose we have the data:

```
x <- c(1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.90, 2.08)

shapiro.test(x)  ##Testing for normality of the data

#####################################################
Shapiro-Wilk normality test

data:  x
W = 0.91699, p-value = 0.4059
#####################################################


x.bar=mean(x)
n=length(x)


##Hyper-parameters
delta=0.14; s.0=1.25; alpha=2.3; beta=1.78


##Parameters of the posterior distributions

gamma=((n*x.bar)+(s.0*delta))/(n+s.0); b.2=var(x)/(n+s.0)
beta.star=0.5*(sum(x^2))-0.5*(delta^2/s.0)-beta+(((n*x.bar)+(s.0*delta))^2)/(2*(n+s.0))


M=10000
install.package("invgamma")  ##For drawing samples from inverse gamma distribution
library(invgamma)

draw=0

for(i in 1:M)
{
  sigma.j=rinvgamma(n=1,shape=alpha+(n/2),shape=beta.star)
  mu.j=rnorm(n=1,mean=gamma,sd=sqrt((sigma.j)/(n+s.0)))
  draw[i]=rnorm(n=1,mean=mu.j,sd=sqrt(sigma.j))
}

k.alpha=quantile(draw,0.05)
k.alpha

      5%
1.421776
```

So we have obtained the k(α) value for the HPD credible set and hence the HPD credible set for the problem is

$$C=\{\mu \in \Re : \pi(\mu|\mathbf{x}) \geq 1.421776 \}$$

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Analysis of Multiparamter Models

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Multinomial Model for Categorical Data

The multinomial sampling distribution is used to describe data for which each observation is one of k possible outcomes.

- If y is the vector of counts of number of observations of each outcome, then

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

where $\sum_{j=1}^{k} \theta_j = 1$ and $\sum_{j=1}^{k} y_j = n$

- The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet distribution.

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

where $0 < \theta_j < 1 \;\; \forall j$ with $\sum_{j=1}^{k} \theta_j = 1$.

- The resulting posterior distribution for the $\theta_j' s$ is a Dirichlet with parameter $\alpha_j + y_j$

- The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet distribution.

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

where $0 < \theta_j < 1 \ \ \forall j$ with $\sum_{j=1}^{k} \theta_j = 1$.

- The resulting posterior distribution for the $\theta_j' s$ is a Dirichlet with parameter $\alpha_j + y_j$

- The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k} \alpha_j$ observations of j^{th} outcome category.
- As in the binomial there are several plausible noninformative Dirichlet prior distributions.
- A uniform density is obtained by setting $\alpha_j = 1$ for all j; this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$.

- The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k} \alpha_j$ observations of j^{th} outcome category.

- As in the binomial there are several plausible noninformative Dirichlet prior distributions.

- A uniform density is obtained by setting $\alpha_j = 1$ for all j; this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$.

- The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k} \alpha_j$ observations of j^{th} outcome category.

- As in the binomial there are several plausible noninformative Dirichlet prior distributions.

- A uniform density is obtained by setting $\alpha_j = 1$ for all j; this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$.

- Setting $\alpha_j = 0$ for all j results an improper prior distribution that is uniform in the $\log(\theta_j)$'s.
- The resulting posterior distribution is proper if there is at least one observation in each of the k categories, so that each component of y is positive.

- Setting $\alpha_j = 0$ for all j results an improper prior distribution that is uniform in the $\log(\theta_j)$'s.

- The resulting posterior distribution is proper if there is at least one observation in each of the k categories, so that each component of y is positive.

Pre-election Polling

For a simple example of a multinomial model, we consider a sample survey question with four possible responses. In late January 2014, a survey was conducted by CSDA of 18591 adults to find out their preferances in upcoming general election of India. Out of 18591 persons , $y_1 = 5205$ supported UPA, $y_2 = 6507$ supported NDA, $y_3 = 929$ supported left and $y_4 = 5950$ supported others

Data:

	Support
UPA	$y_1 = 5205$
NDA	$y_2 = 6507$
Left	$y_3 = 929$
Others	$y_4 = 5950$
Total	$n = 18591$

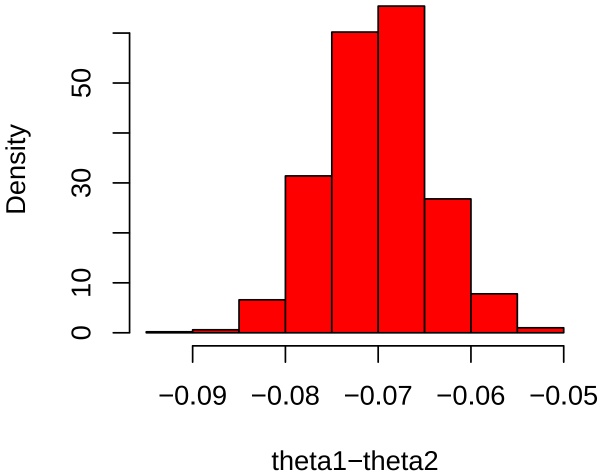
Application Multinomial Distribution

- With a noninformative uniform prior distribution on $\theta, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$
- The posterior distribution for $(\theta_1, \theta_2, \theta_3, \theta_4)$ is Dirichlet(5206,6508,930,5951).
- We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3, \theta_4)$ from posterior Dirichlet distribution and then compute $\theta_1 - \theta_2$ for each.

- With a noninformative uniform prior distribution on $\theta, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$
- The posterior distribution for $(\theta_1, \theta_2, \theta_3, \theta_4)$ is Dirichlet(5206,6508,930,5951).
- We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3, \theta_4)$ from posterior Dirichlet distribution and then compute $\theta_1 - \theta_2$ for each.

- With a noninformative uniform prior distribution on $\theta, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$
- The posterior distribution for $(\theta_1, \theta_2, \theta_3, \theta_4)$ is Dirichlet(5206,6508,930,5951).
- We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3, \theta_4)$ from posterior Dirichlet distribution and then compute $\theta_1 - \theta_2$ for each.

Application Multinomial Distribution



```
   UPA    NDA   Left Others
0.2800 0.3500 0.0501 0.3200

         UPA    NDA   Left Others
2.5%  0.2736 0.3431 0.0469 0.3135
97.5% 0.2865 0.3566 0.0533 0.3266

theta_1 - theta_2:

[1] -0.07

   2.5%   97.5%
-0.0811 -0.0582
```

Application Multinomial Distribution

- The estimated posterior probability NDA had more support than UPA in the survey population is 99%.

- In complicated problems-for example, analyzing the results of many survey questions simultaneously-the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model.

- Formally, additional information can enter the analysis through the prior distribution or the sampling model.

Application Multinomial Distribution

- The estimated posterior probability NDA had more support than UPA in the survey population is 99%.

- In complicated problems-for example, analyzing the results of many survey questions simultaneously-the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model.

- Formally, additional information can enter the analysis through the prior distribution or the sampling model.

Application Multinomial Distribution

- The estimated posterior probability NDA had more support than UPA in the survey population is 99%.

- In complicated problems-for example, analyzing the results of many survey questions simultaneously-the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model.

- Formally, additional information can enter the analysis through the prior distribution or the sampling model.

Conjugate inverse-Wishart family of prior distributions

- Multivariate normal likelihood

$$y|\mu, \Sigma \sim N(\mu, \Sigma)$$

where μ is a d-dimensional column vector and Σ is $d \times d$ symmetric positive-definite covariance matrix

- The likelihood function for a single observation is

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right),$$

Conjugate inverse-Wishart family of prior distributions

- Multivariate normal likelihood

$$y|\mu, \Sigma \sim N(\mu, \Sigma)$$

where μ is a d-dimensional column vector and Σ is $d \times d$ symmetric positive-definite covariance matrix

- The likelihood function for a single observation is

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right),$$

Conjugate inverse-Wishart family of prior distributions

- The likelihood function for a n independent and identically distributed observation is

$$p(y_1, ..., y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)$$
$$= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}tr(\Sigma^{-1}S_0)\right)$$

where S_0 is the matrix of 'sum of squares' relative to μ

$$S_0 = \sum_{i=1}^{n}(y_i-\mu)^T(y_i-\mu)$$

Multivariate Normal with unknown Mean and Covariance

Conjugate inverse-Wishart family of prior distributions

- $\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$
- $\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$
- The joint prior density

$$p(\mu, \Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left\{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\}.$$

Conjugate inverse-Wishart family of prior distributions

- $\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$

- $\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$

- The joint prior density

$$p(\mu, \Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left\{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\}.$$

Conjugate inverse-Wishart family of prior distributions

- $\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$

- $\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$

- The joint prior density

$$
\begin{aligned}
p(\mu, \Sigma) \quad &\propto \quad |\Sigma|^{-((\nu_0+d)/2+1)} \\
&\exp\left\{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\}.
\end{aligned}
$$

Conjugate inverse-Wishart family of prior distributions

- The parameter ν_0 and Λ_0 describe the degrees of freedom and the scale matrix for the inverse-Wishart distribution on Σ.
- The remaining parameters are the prior mean, μ_0, and the number of prior measurements κ_0 on the Σ scale.

Conjugate inverse-Wishart family of prior distributions

- The parameter ν_0 and Λ_0 describe the degrees of freedom and the scale matrix for the inverse-Wishart distribution on Σ.

- The remaining parameters are the prior mean, μ_0, and the number of prior measurements κ_0 on the Σ scale.

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma|y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$

- $\mu_n|\Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$

-
  - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
  - $\kappa_n = \kappa_0 + n$
  - $\nu_n = \nu_0 + n$
  - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$
    where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

Multivariate Normal with unknown Mean and Covariance

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma|y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$
- $\mu_n|\Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$
-
  - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
  - $\kappa_n = \kappa_0 + n$
  - $\nu_n = \nu_0 + n$
  - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$
    where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

Multivariate Normal with unknown Mean and Covariance

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma | y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$
- $\mu_n | \Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$
-
    - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
    - $\kappa_n = \kappa_0 + n$
    - $\nu_n = \nu_0 + n$
    - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$
      where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma|y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$
- $\mu_n|\Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$
-
  - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
  - $\kappa_n = \kappa_0 + n$
  - $\nu_n = \nu_0 + n$
  - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$
    where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

epathshala पाठशाला
MHRD Govt. of India
सत्यमेव जयते

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma | y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$
- $\mu_n | \Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$
-
  - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
  - $\kappa_n = \kappa_0 + n$
  - $\nu_n = \nu_0 + n$
  - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T$
    where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

Conjugate inverse-Wishart family of posterior distributions

- $\Sigma|y_1, ..., y_n \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$

- $\mu_n|\Sigma, y_1, ..., y_n \sim N(\mu_n, \Sigma/\kappa_n)$

-
  - $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$
  - $\kappa_n = \kappa_0 + n$
  - $\nu_n = \nu_0 + n$
  - $\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$
    where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$

- In the development of drugs and other chemical compounds, acute toxicity tests or bioassay experiments are commonly performed on animals.

- Such experiments proceed by administering various dose levels of the compound to batches of animals.

Dose x_i $(\log g/ml)$	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Modeling the dose-response relation

- $$y_i \sim Bin(n_i, \theta_i)$$

  where θ_i is the probability of death for animals given dose x_i.

- The relation ship can be modeled as

  $$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i$$

  this is called logistic regression model.

- It can also be represented as

  $$\theta_i = \frac{\exp\{\alpha + \beta x_i\}}{1 + \exp\{\alpha + \beta x_i\}} = logit^{-1}(\alpha + \beta x_i)$$

Modeling the dose-response relation

- $$y_i \sim Bin(n_i, \theta_i)$$

  where θ_i is the probability of death for animals given dose x_i.

- The relation ship can be modeled as

  $$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i$$

  this is called logistic regression model.

- It can also be represented as

  $$\theta_i = \frac{\exp\{\alpha + \beta x_i\}}{1 + \exp\{\alpha + \beta x_i\}} = logit^{-1}(\alpha + \beta x_i)$$

Modeling the dose-response relation

- $$y_i \sim Bin(n_i, \theta_i)$$

  where θ_i is the probability of death for animals given dose x_i.

- The relation ship can be modeled as

  $$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i$$

  this is called logistic regression model.

- It can also be represented as

  $$\theta_i = \frac{\exp\{\alpha + \beta x_i\}}{1 + \exp\{\alpha + \beta x_i\}} = logit^{-1}(\alpha + \beta x_i)$$

Modeling the dose-response relation

- The likelihood function is

$$p(\alpha, \beta | y, n, x) \propto \prod_{i=1}^{k} [logit^{-1}(\alpha + \beta x_i)]^{y_i} [1 - logit^{-1}(\alpha + \beta x_i)]^{n_i - y_i}$$

- The prior is $p(\alpha, \beta) \sim N(0, \mathbf{I}_2)$

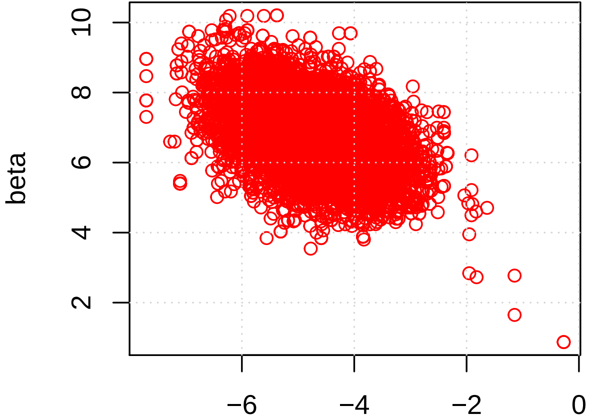
Modeling the dose-response relation

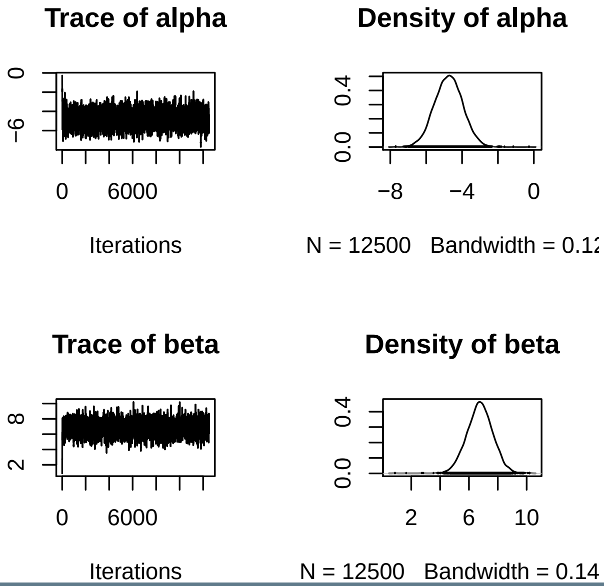
- The likelihood function is

$$p(\alpha,\beta|y,n,x) \propto \prod_{i=1}^{k}[logit^{-1}(\alpha+\beta x_i)]^{y_i}[1-logit^{-1}(\alpha+\beta x_i)]^{n_i-y_i}$$

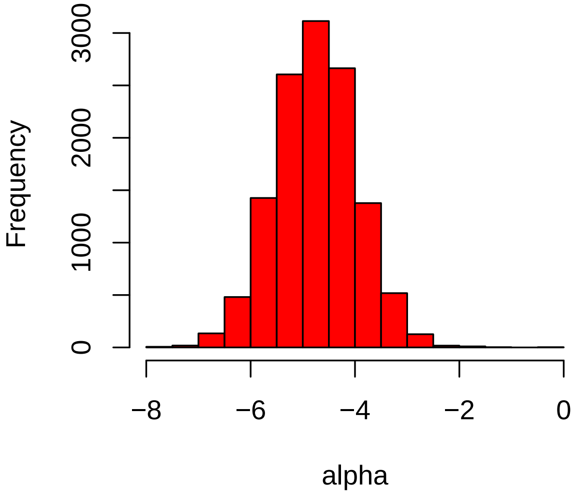
- The prior is $p(\alpha,\beta) \sim N(0,\mathbf{I}_2)$

Simulated samples from posterior distribution

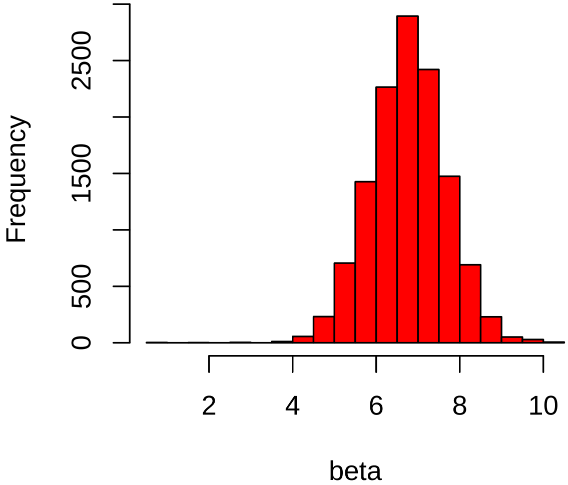



Trace of alpha
0
−6
0
6000
Iterations
Density of alpha
0.4
0.0
−8
−4
0
N = 12500 Bandwidth = 0.12
Trace of beta
8
2
0
6000
Iterations
Density of beta
0.4
0.0
2
6
10
N = 12500 Bandwidth = 0.14

Analysis of a Bioassay Experiment



Analysis of a Bioassay Experiment



A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Introduction to Hierarchical Models

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

- When there are few parameters, posterior inference in nonconjugate multiparameter models can be obtained by simulation methods.

- sophisticated models can often be represented in a hierarchical for which effective computational strategies are available

- When there are few parameters, posterior inference in nonconjugate multiparameter models can be obtained by simulation methods.

- sophisticated models can often be represented in a hierarchical for which effective computational strategies are available

1. Write the likelihood part of the model, $p(y|\theta)$, ignoring any factors that are free of θ.

2. Write the posterior density, $p(\theta|y) \propto p(\theta)p(y|\theta)$. If prior information is well-formulated, include it in $p(\theta)$. Otherwise use non-informative prior

3. Create a crude estimate of the parameters, θ, for use as a starting point and a comparison to the computation in the next step.

4. Draw simulations from $\theta^1, ..., \theta^S$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of θ that may be of interest. For non-conjugate models this step could be difficult.

1. Write the likelihood part of the model, $p(y|\theta)$, ignoring any factors that are free of θ.

2. Write the posterior density, $p(\theta|y) \propto p(\theta)p(y|\theta)$. If prior information is well-formulated, include it in $p(\theta)$. Otherwise use non-informative prior

3. Create a crude estimate of the parameters, θ, for use as a starting point and a comparison to the computation in the next step.

4. Draw simulations from $\theta^1, ..., \theta^S$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of θ that may be of interest. For non-conjugate models this step could be difficult.

1. Write the likelihood part of the model, $p(y|\theta)$, ignoring any factors that are free of θ.

2. Write the posterior density, $p(\theta|y) \propto p(\theta)p(y|\theta)$. If prior information is well-formulated, include it in $p(\theta)$. Otherwise use non-informative prior

3. Create a crude estimate of the parameters, θ, for use as a starting point and a comparison to the computation in the next step.

4. Draw simulations from $\theta^1, ..., \theta^S$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of θ that may be of interest. For non-conjugate models this step could be difficult.

1. Write the likelihood part of the model, $p(y|\theta)$, ignoring any factors that are free of θ.

2. Write the posterior density, $p(\theta|y) \propto p(\theta)p(y|\theta)$. If prior information is well-formulated, include it in $p(\theta)$. Otherwise use non-informative prior

3. Create a crude estimate of the parameters, θ, for use as a starting point and a comparison to the computation in the next step.

4. Draw simulations from $\theta^1, ..., \theta^S$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of θ that may be of interest. For non-conjugate models this step could be difficult.

5 If any predictive quantities, $\tilde{y}$, are of interest simulate $\tilde{y}^1, ..., \tilde{y}^S$ by drawing each $\tilde{y}^s$ from the sampling distribution conditional on the drawn value θ^s, $p(\tilde{y}|\theta^s)$.

- Various methods (such as Markov Chain Monte Carlo) have been developed to draw posterior simulations in complicated models.
- If θ has only one or two components, it is possible to draw simulations by computing on a grid.

5 If any predictive quantities, $\tilde{y}$, are of interest simulate $\tilde{y}^1, ..., \tilde{y}^S$ by drawing each $\tilde{y}^s$ from the sampling distribution conditional on the drawn value θ^s, $p(\tilde{y}|\theta^s)$.

- Various methods (such as Markov Chain Monte Carlo) have been developed to draw posterior simulations in complicated models.

- If θ has only one or two components, it is possible to draw simulations by computing on a grid.

5 If any predictive quantities, $\tilde{y}$, are of interest simulate $\tilde{y}^1, ..., \tilde{y}^S$ by drawing each $\tilde{y}^s$ from the sampling distribution conditional on the drawn value θ^s, $p(\tilde{y}|\theta^s)$.

- Various methods (such as Markov Chain Monte Carlo) have been developed to draw posterior simulations in complicated models.
- If θ has only one or two components, it is possible to draw simulations by computing on a grid.

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
  - e.g. we collect measurements of individuals who live in a certain locality or belong to a particular race or social group.
- When this occurs, standard techniques either assume that these groups belong to entirely different populations or ignore the aggregate information entirely.
- Hierarchical models provide a way of pooling the information for the disparate groups without assuming that they belong to precisely the same population.

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
  - e.g. we collect measurements of individuals who live in a certain locality or belong to a particular race or social group.
- When this occurs, standard techniques either assume that these groups belong to entirely different populations or ignore the aggregate information entirely.
- Hierarchical models provide a way of pooling the information for the disparate groups without assuming that they belong to precisely the same population.

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
  - e.g. we collect measurements of individuals who live in a certain locality or belong to a particular race or social group.
- When this occurs, standard techniques either assume that these groups belong to entirely different populations or ignore the aggregate information entirely.
- Hierarchical models provide a way of pooling the information for the disparate groups without assuming that they belong to precisely the same population.

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
  - e.g. we collect measurements of individuals who live in a certain locality or belong to a particular race or social group.
- When this occurs, standard techniques either assume that these groups belong to entirely different populations or ignore the aggregate information entirely.
- Hierarchical models provide a way of pooling the information for the disparate groups without assuming that they belong to precisely the same population.

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.

- Let Y_{ij} represent observation j from population i

- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.

- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.

- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

Hierarchical Models

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.
- Let Y_{ij} represent observation j from population i
- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.
- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.
- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.
- Let Y_{ij} represent observation j from population i
- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.
- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.
- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.

- Let Y_{ij} represent observation j from population i

- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.

- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.

- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.
- Let Y_{ij} represent observation j from population i
- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.
- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.
- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

- Now we extend the model, and assume that the parameters Θ_{11}, Θ_{12} that govern the distribution of the Θ's are themselves random variables and assign a prior distribution to these variables as well:

$$\Theta \sim f(\alpha, \beta)$$

- Θ is called hyperprior. The parameters a, b, c, d for the hyperprior may be "known" and represent our prior beliefs about Θ.

- In theory, we can also assign a probability distribution for these quantities as well, and proceed to another layer of hierarchy.

ePathshala पाठशाला MHRD Govt. of India

- Now we extend the model, and assume that the parameters Θ_{11}, Θ_{12} that govern the distribution of the Θ's are themselves random variables and assign a prior distribution to these variables as well:

$$\Theta \sim f(\alpha, \beta)$$

- Θ is called hyperprior. The parameters a, b, c, d for the hyperprior may be "known" and represent our prior beliefs about Θ.

- In theory, we can also assign a probability distribution for these quantities as well, and proceed to another layer of hierarchy.

- Now we extend the model, and assume that the parameters Θ_{11}, Θ_{12} that govern the distribution of the Θ's are themselves random variables and assign a prior distribution to these variables as well:

$$\Theta \sim f(\alpha, \beta)$$

- Θ is called hyperprior. The parameters a, b, c, d for the hyperprior may be "known" and represent our prior beliefs about Θ.

- In theory, we can also assign a probability distribution for these quantities as well, and proceed to another layer of hierarchy.

$$
\begin{array}{ccccccccc}
 & & & & \alpha, \beta & & & & \\
\theta_1 & \theta_2 & \theta_3 & \cdots & \cdots & \cdots & \cdots & \theta_{70} & \theta_{71} \\
\downarrow & \downarrow & \downarrow & & & & & \downarrow & \downarrow \\
y_1 & y_2 & y_3 & \cdots & \cdots & \cdots & \cdots & y_{70} & y_{71}
\end{array}
$$

Exchangeability : Formal

The parameters $\theta_1, \theta_2, ..., \theta_n$ are exchangeable in their joint distribution if $p(\theta_1, \theta_2, ..., \theta_n)$ is invariant to permutations in the index $1, 2, ..., n$.

Exchangeability : Informal

If no information other than the data is available to distinguish any of the θ_j's form any of the others, and no ordering of the parameters can be made, one must assume symmetry among the parameters in the prior distribution.

This concept is closely related to the concept of identically and independent random variables where, conditional on the data, each observation is treated the same.

Exchangeability : Formal

The parameters $\theta_1, \theta_2, ..., \theta_n$ are exchangeable in their joint distribution if $p(\theta_1, \theta_2, ..., \theta_n)$ is invariant to permutations in the index $1, 2, ..., n$.

Exchangeability : Informal

If no information other than the data is available to distinguish any of the θ_j's form any of the others, and no ordering of the parameters can be made, one must assume symmetry among the parameters in the prior distribution.

This concept is closely related to the concept of identically and independent random variables where, conditional on the data, each observation is treated the same.

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?

- We also have the times (t_i) at which each pump was observed.

- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

Application: Poisson-Gamma Model

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?
- We also have the times (t_i) at which each pump was observed.
- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

Application: Poisson-Gamma Model

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?
- We also have the times (t_i) at which each pump was observed.
- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

Application: Poisson-Gamma Model

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?

- We also have the times (t_i) at which each pump was observed.

- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

Application: Poisson-Gamma Model

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?
- We also have the times (t_i) at which each pump was observed.
- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

Exchangeability

- Exchangeability means that we can treat the parameters for each sub-population as exchangeable units.

- In its simplest form, each parameter θ_j is treated as an independent sample from a distribution governed by unknown parameter vector Θ.

$$p(\theta_1, \theta_2, ..., \theta_n|\Theta) = \prod_i p(\theta_i|\Theta)$$

- In a more general form, we may also condition on data that we have about the different sub-populations.

- Exchangeability means that we can treat the parameters for each sub-population as exchangeable units.

- In its simplest form, each parameter θ_j is treated as an independent sample from a distribution governed by unknown parameter vector Θ.

$$p(\theta_1, \theta_2, ..., \theta_n|\Theta) = \prod_i p(\theta_i|\Theta)$$

- In a more general form, we may also condition on data that we have about the different sub-populations.

Exchangeability

- Exchangeability means that we can treat the parameters for each sub-population as exchangeable units.

- In its simplest form, each parameter θ_j is treated as an independent sample from a distribution governed by unknown parameter vector Θ.

$$p(\theta_1, \theta_2, ..., \theta_n|\Theta) = \prod_i p(\theta_i|\Theta)$$

- In a more general form, we may also condition on data that we have about the different sub-populations.

- We can write the joint prior distribution as:

$$p(\theta_1, \theta_2, ..., \theta_n, \Theta) = p(\theta_1, \theta_2, ..., \theta_n | \Theta) p(\Theta)$$

- By Baye's Rule

$$p(\theta_1, ..., \theta_n, \Theta | Y) \propto \text{prior} \times \text{likelihood for } Y$$

- We can write the joint prior distribution as:

$$p(\theta_1, \theta_2, ..., \theta_n, \Theta) = p(\theta_1, \theta_2, ..., \theta_n|\Theta)p(\Theta)$$

- By Baye's Rule

$$p(\theta_1, ..., \theta_n, \Theta|Y) \propto \text{prior} \times \text{likelihood for } Y$$

Application: Poisson-Gamma Model

- We consider the data model to be:

$$y_i \sim Poisson(\lambda_i t_i) \quad \text{for } i = 1, ..., 10.$$

- To model this as a hierarchical process, we assume that each of the absence λ_i are exchangeable draws from a common distribution.

- In this case, the gamma distribution has desirable properties.

$$\lambda_i \sim Gamma(\alpha, \beta) \quad \text{for } i = 1, ..., 10.$$

Note that $\alpha = 1.8$ and β are unknown parameters.

- We consider the data model to be:

$$y_i \sim Poisson(\lambda_i t_i) \quad \text{for } i = 1, ..., 10.$$

- To model this as a hierarchical process, we assume that each of the absence λ_i are exchangeable draws from a common distribution.

- In this case, the gamma distribution has desirable properties.

$$\lambda_i \sim Gamma(\alpha, \beta) \quad \text{for } i = 1, ..., 10.$$

Note that $\alpha = 1.8$ and β are unknown parameters.

- To satisfy the requirement of exchangeability, what must we assume about the data generating process?

- Finally, to complete the hierarchical structure, we must assign "hyperpriors" for the parameters on β. Again, the gamma distribution has nice properties, so we assume that:

$$\beta \sim Gamma(\nu, \delta)$$

- assign $Gamma(\nu, \delta)$ on unknow β with $\nu = 0.001$ and $\delta = 1$.

- To satisfy the requirement of exchangeability, what must we assume about the data generating process?

- Finally, to complete the hierarchical structure, we must assign "hyperpriors" for the parameters on β. Again, the gamma distribution has nice properties, so we assume that:

$$\beta \sim Gamma(\nu, \delta)$$

- assign $Gamma(\nu, \delta)$ on unknow β with $\nu = 0.001$ and $\delta = 1$.

- To satisfy the requirement of exchangeability, what must we assume about the data generating process?

- Finally, to complete the hierarchical structure, we must assign "hyperpriors" for the parameters on β. Again, the gamma distribution has nice properties, so we assume that:

$$\beta \sim Gamma(\nu, \delta)$$

- assign $Gamma(\nu, \delta)$ on unknow β with $\nu = 0.001$ and $\delta = 1$.

Application: Poisson-Gamma Model

- The joint posterior distribution is:

$$p(\lambda_i, \beta|y,t) \propto \prod_i Pois(\lambda_i t_i|y_i) \times Gamma(\lambda_i|\alpha,\beta) Gamma(\beta|\nu,\delta)$$

- Using our trick for conditional distributions, we know

$$p(\lambda_i|\beta,y,t) \sim Gamma(y_i+\alpha, t_i+\beta)$$

- $p(\beta|\lambda_1,y,t) \sim Gamma(10\lambda+\nu, \delta+\sum_{i=1}^{1} 0\lambda_i)$

Application: Poisson-Gamma Model

- The joint posterior distribution is:

$$p(\lambda_i, \beta|y, t) \propto \prod_i Pois(\lambda_i t_i|y_i) \times Gamma(\lambda_i|\alpha, \beta) Gamma(\beta|\nu, \delta)$$

- Using our trick for conditional distributions, we know

$$p(\lambda_i|\beta, y, t) \sim Gamma(y_i + \alpha, t_i + \beta)$$

- $p(\beta|\lambda_1, y, t) \sim Gamma(10\lambda + \nu, \delta + \sum_{i=1}^{1} 0\lambda_i)$

- The joint posterior distribution is:

$$p(\lambda_i, \beta | y, t) \propto \prod_i Pois(\lambda_i t_i | y_i) \times Gamma(\lambda_i | \alpha, \beta) Gamma(\beta | \nu, \delta)$$

- Using our trick for conditional distributions, we know

$$p(\lambda_i | \beta, y, t) \sim Gamma(y_i + \alpha, t_i + \beta)$$

- $p(\beta | \lambda_1, y, t) \sim Gamma(10\lambda + \nu, \delta + \sum_{i=1}^{1} 0\lambda_i)$

Application: Poisson-Gamma Model

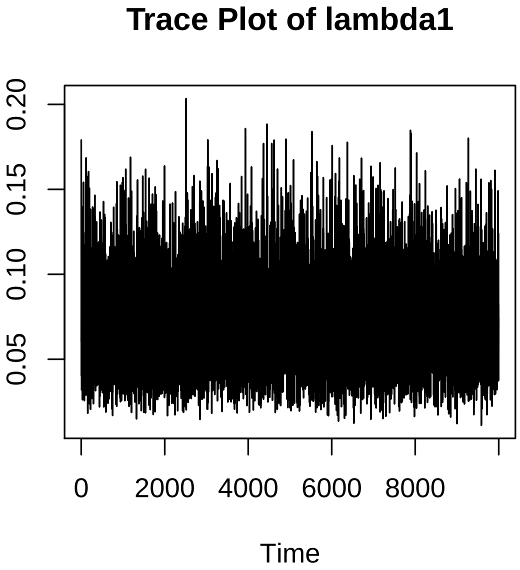
```
Posterior mean of lambda's :

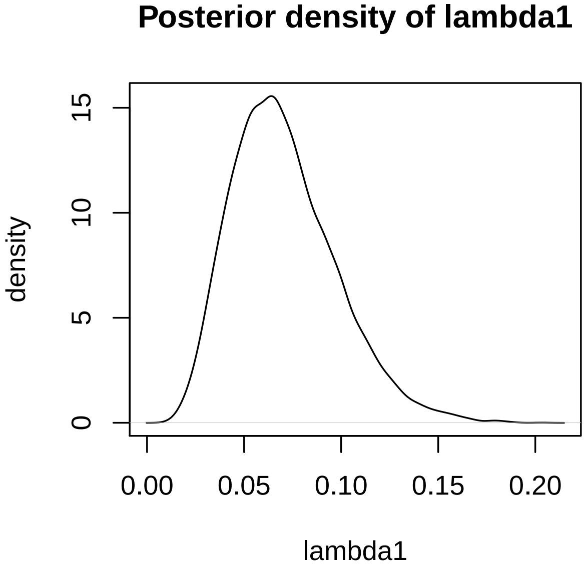
 [1] 0.07063148 0.15167434 0.10484798 0.12280676 0.65433702
 [7] 0.84803853 0.86260141 1.36635800 1.92780912

Posterior mean of beta :

[1] 2.393286
```

Application: Poisson-Gamma Model




Posterior density of lambda1
density
0
5
10
15
0.00
0.05
0.10
0.15
0.20
lambda1

- As we assume $\lambda_i \sim Gamma(\alpha = 1.8, \beta)$ and estimated $\hat{\beta} = 2.39$
- 95% CI of $Gamma(\alpha = 1.8, \hat{\beta} = 2.39)$
  `[1] 0.07631031 2.17514481`
- It indicates λ_1 might be an outlier

- As we assume $\lambda_i \sim Gamma(\alpha = 1.8, \beta)$ and estimated $\hat{\beta} = 2.39$
- 95% CI of $Gamma(\alpha = 1.8, \hat{\beta} = 2.39)$

```
[1] 0.07631031 2.17514481
```

- It indicates λ_1 might be an outlier

Application: Poisson-Gamma Model

- As we assume $\lambda_i \sim Gamma(\alpha = 1.8, \beta)$ and estimated $\hat{\beta} = 2.39$
- 95% CI of $Gamma(\alpha = 1.8, \hat{\beta} = 2.39)$

```
[1] 0.07631031 2.17514481
```

- It indicates λ_1 might be an outlier

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Monte Carlo Intergration and Simulation Technique - Part 1

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y)d\theta = g(y)$$

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y)d\theta = g(y)$$

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_\Theta \theta p(\theta|y)d\theta = g(y)$$

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_\Theta \theta p(\theta|y)d\theta = g(y)$$

- In order to get the posterior mean we have to solve this integration
- Analytical solution does not exists for many sophisticated models
- So we have to resort to simulation technique to solve this integration problem.
- Typically it is known as Monte Carlo Integration method

Introduction

- In order to get the posterior mean we have to solve this integration
- Analytical solution does not exists for many sophisticated models
- So we have to resort to simulation technique to solve this integration problem.
- Typically it is known as Monte Carlo Integration method

- In order to get the posterior mean we have to solve this integration
- Analytical solution does not exists for many sophisticated models
- So we have to resort to simulation technique to solve this integration problem.
- Typically it is known as Monte Carlo Integration method

- In order to get the posterior mean we have to solve this integration
- Analytical solution does not exists for many sophisticated models
- So we have to resort to simulation technique to solve this integration problem.
- Typically it is known as Monte Carlo Integration method

- Monte Carlo methods rely on
  - The possibility of generation of endless flow of random variables
  - For well-known or new distributions.
- Such a simulation is based on the generation of uniform random variables on the interval $(0, 1)$.
- We are not concerned with the details of producing uniform random numbers.
- We assume the existence of such a sequence

- Monte Carlo methods rely on
  - The possibility of generation of endless flow of random variables
  - For well-known or new distributions.
- Such a simulation is based on the generation of uniform random variables on the interval $(0, 1)$.
- We are not concerned with the details of producing uniform random numbers.
- We assume the existence of such a sequence

e-Pathshala पाठशाला
MHRD Govt. of India
सत्यमेव जयते

- Monte Carlo methods rely on
  - The possibility of generation of endless flow of random variables
  - For well-known or new distributions.
- Such a simulation is based on the generation of uniform random variables on the interval $(0,1)$.
- We are not concerned with the details of producing uniform random numbers.
- We assume the existence of such a sequence

Monte Carlo Methods

- Monte Carlo methods rely on
  - The possibility of generation of endless flow of random variables
  - For well-known or new distributions.
- Such a simulation is based on the generation of uniform random variables on the interval $(0, 1)$.
- We are not concerned with the details of producing uniform random numbers.
- We assume the existence of such a sequence

- Monte Carlo methods rely on
  - The possibility of generation of endless flow of random variables
  - For well-known or new distributions.
- Such a simulation is based on the generation of uniform random variables on the interval $(0, 1)$.
- We are not concerned with the details of producing uniform random numbers.
- We assume the existence of such a sequence

- As we want to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta = g(y)$$

we can do so by simulating random samples from $p(\theta|y)$

- Suppose $(\theta^1, ..., \theta^N)$ are random samples from $p(\theta|y)$, then we can approximate $g(y)$ by

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s$$

- If we can ensure simple random sample then SLLN ensures

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s \to g(y) = E(\theta|y)$$

as $N \to \infty$, where N is the simulation size.

- As we want to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta = g(y)$$

we can do so by simulating random samples from $p(\theta|y)$

- Suppose $(\theta^1, ..., \theta^N)$ are random samples from $p(\theta|y)$, then we can approximate $g(y)$ by

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s$$

- If we can ensure simple random sample then SLLN ensures

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s \to g(y) = E(\theta|y)$$

as $N \to \infty$, where N is the simulation size.

Monte Carlo Integration

- As we want to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y)d\theta = g(y)$$

we can do so by simulating random samples from $p(\theta|y)$

- Suppose $(\theta^1, ..., \theta^N)$ are random samples from $p(\theta|y)$, then we can approximate $g(y)$ by

$$\hat{g}(y) = \frac{1}{N}\sum_{s=1}^{N}\theta^s$$

- If we can ensure simple random sample then SLLN ensures

$$\hat{g}(y) = \frac{1}{N}\sum_{s=1}^{N}\theta^s \rightarrow g(y) = E(\theta|y)$$

as $N \rightarrow \infty$, where N is the simulation size.

- As we want to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta = g(y)$$

we can do so by simulating random samples from $p(\theta|y)$

- Suppose $(\theta^1, ..., \theta^N)$ are random samples from $p(\theta|y)$, then we can approximate $g(y)$ by

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s$$

- If we can ensure simple random sample then SLLN ensures

$$\hat{g}(y) = \frac{1}{N} \sum_{s=1}^{N} \theta^s \rightarrow g(y) = E(\theta|y)$$

as $N \rightarrow \infty$, where N is the simulation size.

Using the R Generator

- R has a large number of functions that will generate random samples from standard distributions.
- If built-in function in R is available then we can use it directly.
- However, if the built-in functions are not available and the posterior distribution is known upto its kernel then we have to use genric methods to draw samples from such distributions.

Using the R Generator

- R has a large number of functions that will generate random samples from standard distributions.

- If built-in function in R is available then we can use it directly.

- However, if the built-in functions are not available and the posterior distribution is known upto its kernel then we have to use genric methods to draw samples from such distributions.

Using the R Generator

- R has a large number of functions that will generate random samples from standard distributions.
- If built-in function in R is available then we can use it directly.
- However, if the built-in functions are not available and the posterior distribution is known upto its kernel then we have to use genric methods to draw samples from such distributions.

e-Pathshala पाठशाला MHRD Govt. of India

- The probability integral transform allows us to transform a uniform into any random variable.

- If X has a density f and cdf F, then we have the relation

$$F(x) = \int_{-\infty}^{x} f(u)du$$

  we know $F(x) \sim Unif(0, 1)$ - so we set $U = F(x)$ and solve for x

- If $X \sim Exp(1)$ then $F(x) = 1 - e^{-x}$

- solving for x in $u = 1 - e^{-x}$ gives $x = -\log(1 - u)$

- The probability integral transform allows us to transform a uniform into any random variable.

- If X has a density f and cdf F, then we have the relation

$$F(x) = \int_{-\infty}^{x} f(u)du$$

  we know $F(x) \sim Unif(0,1)$ - so we set $U = F(x)$ and solve for x

- If $X \sim Exp(1)$ then $F(x) = 1 - e^{-x}$

- solving for x in $u = 1 - e^{-x}$ gives $x = -\log(1-u)$

- The probability integral transform allows us to transform a uniform into any random variable.

- If X has a density f and cdf F, then we have the relation

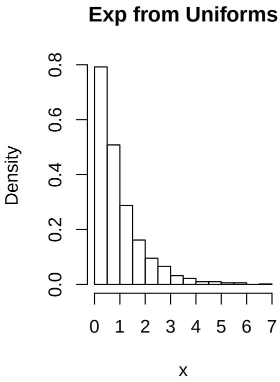
$$F(x) = \int_{-\infty}^{x} f(u)du$$

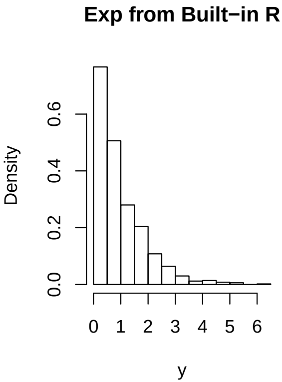
  we know $F(x) \sim Unif(0,1)$ - so we set $U = F(x)$ and solve for x

- If $X \sim Exp(1)$ then $F(x) = 1 - e^{-x}$

- solving for x in $u = 1 - e^{-x}$ gives $x = -\log(1-u)$

Generating Exponentially Distributed Random Samples





The Inverse Transform from Uniforms

- This method is useful for other probability distributions ; ones obtained as a transformation of uniform random variables

- Logistic cdf:

$$F(x) = \frac{1}{1 + \exp\left(-\frac{x-\mu}{\lambda}\right)}$$

- Cauchy cdf:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctan((x - \mu)/\sigma)$$

- This method is useful for other probability distributions ; ones obtained as a transformation of uniform random variables

- Logistic cdf:

$$F(x) = \frac{1}{1 + \exp\left(-\frac{x-\mu}{\lambda}\right)}$$

- Cauchy cdf:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctan((x-\mu)/\sigma)$$

General Transform Method

- If $X_i \overset{iid}{\sim} Exp(1)$; three standard distribution can be derived as
  - $Y = 2\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$
  - $Y = \beta\sum_{i=1}^{n} X_i \sim Gamma(n, \beta)$
  - $Y = \frac{\sum_{i=1}^{n_1} X_i}{\sum_{i=1}^{n_1+n_2} X_i} \sim Beta(n_1, n_2)$

  where $n \in \mathcal{N} = \{1, 2, ...\}$
- These transformations are quite simple and we will use them quite often
- However, there is a limit to their usefullness. Only when closed form CDF is available the method is available.
- The method will not work for Gaussian distribution as the CDF is not in closed form.

- If $X_i \stackrel{iid}{\sim} Exp(1)$; three standard distribution can be derived as
  - $Y = 2\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$
  - $Y = \beta\sum_{i=1}^{n} X_i \sim Gamma(n, \beta)$
  - $Y = \frac{\sum_{i=1}^{n_1} X_i}{\sum_{i=1}^{n_1+n_2} X_i} \sim Beta(n_1, n_2)$

  where $n \in \mathcal{N} = \{1, 2, ...\}$
- These transformations are quite simple and we will use them quite often
- However, there is a limit to their usefullness. Only when closed form CDF is available the method is available.
- The method will not work for Gaussian distribution as the CDF is not in closed form.

Pathshala पाठशाला MHRD Govt. of India

- If $X_i \overset{iid}{\sim} Exp(1)$; three standard distribution can be derived as
  - $Y = 2\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$
  - $Y = \beta \sum_{i=1}^{n} X_i \sim Gamma(n, \beta)$
  - $Y = \frac{\sum_{i=1}^{n_1} X_i}{\sum_{i=1}^{n_1+n_2} X_i} \sim Beta(n_1, n_2)$

  where $n \in \mathcal{N} = \{1, 2, ...\}$
- These transformations are quite simple and we will use them quite often
- However, there is a limit to their usefullness. Only when closed form CDF is available the method is available.
- The method will not work for Gaussian distribution as the CDF is not in closed form.

General Transform Method

- If $X_i \overset{iid}{\sim} Exp(1)$; three standard distribution can be derived as
  - $Y = 2\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$
  - $Y = \beta\sum_{i=1}^{n} X_i \sim Gamma(n, \beta)$
  - $Y = \frac{\sum_{i=1}^{n_1} X_i}{\sum_{i=1}^{n_1+n_2} X_i} \sim Beta(n_1, n_2)$

  where $n \in \mathcal{N} = \{1, 2, ...\}$
- These transformations are quite simple and we will use them quite often
- However, there is a limit to their usefullness. Only when closed form CDF is available the method is available.
- The method will not work for Gaussian distribution as the CDF is not in closed form.

- Box-Muller transform to generate from Normal distribution
- If U_1 and U_2 are iid from $Unif(0,1)$
- The variable X_1 and X_2

  $$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad \text{and} \quad X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  are iid $N(0,1)$ by virtue of a change of variable argument.
- The Box-Muller algorithm is exact, not a crude CLT-based approximation
- Note that this is not the generator implemented in R. It uses the probability inverse transform with a very accurate representation of the normal cdf

- Box-Muller transform to generate from Normal distribution
- If U_1 and U_2 are iid from $Unif(0,1)$
- The variable X_1 and X_2

  $$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad \text{and} \quad X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  are iid $N(0,1)$ by virtue of a change of variable argument.
- The Box-Muller algorithm is exact, not a crude CLT-based approximation
- Note that this is not the generator implemented in R. It uses the probability inverse transform with a very accurate representation of the normal cdf

- Box-Muller transform to generate from Normal distribution
- If U_1 and U_2 are iid from $Unif(0,1)$
- The variable X_1 and X_2

  $X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$, and $X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$

  are iid $N(0,1)$ by virtue of a change of variable argument.
- The Box-Muller algorithm is exact, not a crude CLT-based approximation
- Note that this is not the generator implemented in R. It uses the probability inverse transform with a very accurate representation of the normal cdf

- Box-Muller transform to generate from Normal distribution
- If U_1 and U_2 are iid from $Unif(0,1)$
- The variable X_1 and X_2

  $$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad \text{and} \quad X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  are iid $N(0,1)$ by virtue of a change of variable argument.
- The Box-Muller algorithm is exact, not a crude CLT-based approximation
- Note that this is not the generator implemented in R. It uses the probability inverse transform with a very accurate representation of the normal cdf

- Box-Muller transform to generate from Normal distribution
- If U_1 and U_2 are iid from $Unif(0,1)$
- The variable X_1 and X_2

  $$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad \text{and} \quad X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  are iid $N(0,1)$ by virtue of a change of variable argument.
- The Box-Muller algorithm is exact, not a crude CLT-based approximation
- Note that this is not the generator implemented in R. It uses the probability inverse transform with a very accurate representation of the normal cdf

Accept-Reject Method

- There are many distributions where transform methods fail
- For these cases, we must turn to indirect methods
  - We generate a candidate random variable
  - Only accept it subject to passing a test
- This class of 'Accept-Reject' methods is extremely powerful. It will allow us to simulate from virtually any distribution.
- Accept-Reject Methods:
  - Only require the functional form of the density f of interest
  - f: target density, g: candidate density

  where it is simpler to simulate random variables from g

Accept-Reject Method

- There are many distributions where transform methods fail
- For these cases, we must turn to indirect methods
  - We generate a candidate random variable
  - Only accept it subject to passing a test
- This class of 'Accept-Reject' methods is extremely powerful. It will allow us to simulate from virtually any distribution.
- Accept-Reject Methods:
  - Only require the functional form of the density f of interest
  - f: target density, g: candidate density

  where it is simpler to simulate random variables from g

Accept-Reject Method

- There are many distributions where transform methods fail
- For these cases, we must turn to indirect methods
  - We generate a candidate random variable
  - Only accept it subject to passing a test
- This class of 'Accept-Reject' methods is extremely powerful. It will allow us to simulate from virtually any distribution.
- Accept-Reject Methods:
  - Only require the functional form of the density f of interest
  - f: target density, g: candidate density

  where it is simpler to simulate random variables from g

Accept-Reject Method

- There are many distributions where transform methods fail
- For these cases, we must turn to indirect methods
  - We generate a candidate random variable
  - Only accept it subject to passing a test
- This class of 'Accept-Reject' methods is extremely powerful. It will allow us to simulate from virtually any distribution.
- Accept-Reject Methods:
  - Only require the functional form of the density f of interest
  - f: target density, g: candidate density

  where it is simpler to simulate random variables from g

Accept-Reject Method

- There are many distributions where transform methods fail
- For these cases, we must turn to indirect methods
  - We generate a candidate random variable
  - Only accept it subject to passing a test
- This class of 'Accept-Reject' methods is extremely powerful. It will allow us to simulate from virtually any distribution.
- Accept-Reject Methods:
  - Only require the functional form of the density f of interest
  - f: target density, g: candidate density

  where it is simpler to simulate random variables from g

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0,1)$
  - If $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

e-pathshala पाठशाला MHRD Govt. of India

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0, 1)$
  - If $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0, 1)$
  - If $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0, 1)$
  - If $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

Accept-Reject Methods

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0, 1)$
  - If $U \leq \frac{1}{M} \frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

- The only constraints we impose on this candidate density g
  - f and g have compatible supports (i.e., $g(x) > 0$ when $f(x) > 0$).
- $X \sim f$ can be simulated as follows:
  - Generate $Y \sim g$ and independently generate $U \sim unif(0, 1)$
  - If $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$ set $X = Y$
  - If inequality is not satisfied then discard Y and U and start again.
- Note $M = \sup_x \frac{f(x)}{g(x)}$
- $P(\text{Accept}) = \frac{1}{M}$ and expected waiting time M.

Accept-Reject Algorithm

1. Generate $Y \sim g$, $U \sim uinf(0, 1)$
2. Accept $X = Y$ if $U \leq f(Y)/Mg(Y)$
3. Return to 1 otherwise

Accept-Reject Methods: Theory

- Why does this method work?

- A straightforward probability calculation shows
$$P(Y \leq x|\text{Accept}) = P\left(Y \leq x|U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$
where $Y \sim g$ and $X \sim f$.

- Simulating from g, the output of this algorithm is exactly distributed from f.

- The Accept-Reject method is applicable in any dimension.

- As long as g is a density over the same space as f.

- Only need to know f/g upto a constant

Accept-Reject Methods: Theory

- Why does this method work?

- A straightforward probability calculation shows
$$P(Y \leq x|\text{Accept}) = P\left(Y \leq x|U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$
where $Y \sim g$ and $X \sim f$.

- Simulating from g, the output of this algorithm is exactly distributed from f.

- The Accept-Reject method is applicable in any dimension.

- As long as g is a density over the same space as f.

- Only need to know f/g upto a constant

Accept-Reject Methods: Theory

- Why does this method work?

- A straightforward probability calculation shows
$$P(Y \leq x | \text{Accept}) = P\left(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$
where $Y \sim g$ and $X \sim f$.

- Simulating from g, the output of this algorithm is exactly distributed from f.

- The Accept-Reject method is applicable in any dimension.

- As long as g is a density over the same space as f.

- Only need to know f/g upto a constant

Accept-Reject Methods: Theory

- Why does this method work?
- A straightforward probability calculation shows
$$P(Y \leq x|\text{Accept}) = P\left(Y \leq x|U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$
where $Y \sim g$ and $X \sim f$.
- Simulating from g, the output of this algorithm is exactly distributed from f.
- The Accept-Reject method is applicable in any dimension.
- As long as g is a density over the same space as f.
- Only need to know f/g upto a constant

Accept-Reject Methods: Theory

- Why does this method work?
- A straightforward probability calculation shows

$$P(Y \leq x|\text{Accept}) = P\left(Y \leq x|U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$

  where $Y \sim g$ and $X \sim f$.
- Simulating from g, the output of this algorithm is exactly distributed from f.
- The Accept-Reject method is applicable in any dimension.
- As long as g is a density over the same space as f.
- Only need to know f/g upto a constant

- Why does this method work?

- A straightforward probability calculation shows

$$P(Y \le x|\text{Accept}) = P\left(Y \le x|U \le \frac{f(Y)}{Mg(Y)}\right) = P(X \le x),$$

  where $Y \sim g$ and $X \sim f$.

- Simulating from g, the output of this algorithm is exactly distributed from f.

- The Accept-Reject method is applicable in any dimension.

- As long as g is a density over the same space as f.

- Only need to know f/g upto a constant

Accept-Reject Methods: Theory

- Why does this method work?

- A straightforward probability calculation shows

$$P(Y \leq x | \text{Accept}) = P\left(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right) = P(X \leq x),$$

  where $Y \sim g$ and $X \sim f$.

- Simulating from g, the output of this algorithm is exactly distributed from f.

- The Accept-Reject method is applicable in any dimension.

- As long as g is a density over the same space as f.

- Only need to know f/g upto a constant

- Generate $X \sim Beta(a, b)$
- No direct method available if a and b are not integers
- Generate for $a = 2.3$ and $b = 7.9$
- We can generate if a and b is integer

- Generate $X \sim Beta(a, b)$
- No direct method available if a and b are not integers
- Generate for $a = 2.3$ and $b = 7.9$
- We can generate if a and b is integer

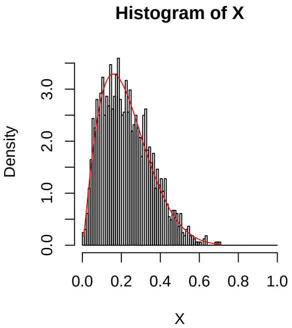
- Generate $X \sim Beta(a, b)$
- No direct method available if a and b are not integers
- Generate for $a = 2.3$ and $b = 7.9$
- We can generate if a and b is integer

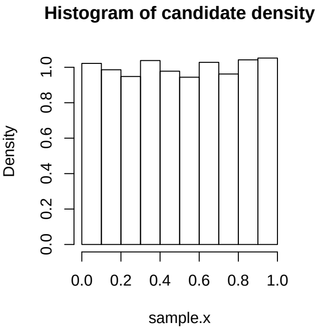
- Generate $X \sim Beta(a, b)$
- No direct method available if a and b are not integers
- Generate for $a = 2.3$ and $b = 7.9$
- We can generate if a and b is integer

Accept-Rejection Algorithm for Beta Distribution

- Candidate distribution $Unif(0, 1)$
- Target distribution $Beta(2.3, 7.9)$

```
Acceptance Rate :
[1] 32.84
```

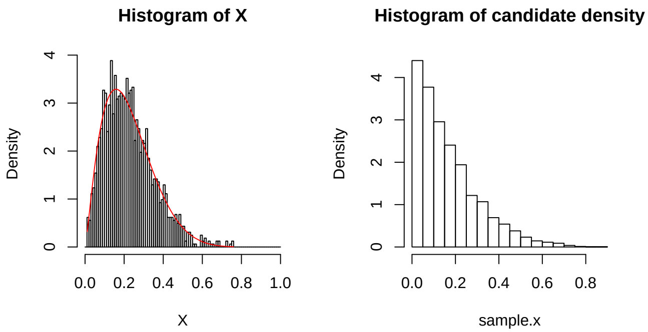




Accept-Rejection Algorithm for Beta Distribution

- Candidate distribution $Beta(1,5)$
- Target distribution $Beta(2.3,7.9)$

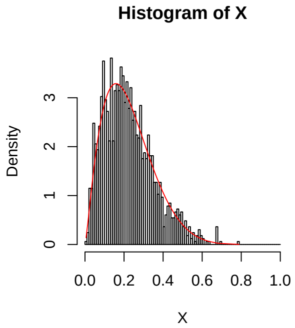
```
Acceptance Rate :
[1] 32.42
```

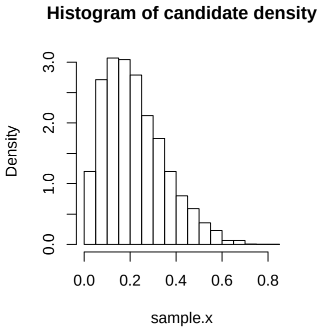


Accept-Rejection Algorithm for Beta Distribution

- Candidate distribution $Beta(2, 7)$
- Target distribution $Beta(2.3, 7.9)$

```
Acceptance Rate :
[1] 33.06
```





A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

MHRD Govt. of India

Subject: Statistics

Paper: Statistical Inference
Module: Monte Carlo Intergration and Simulation Technique - Part 2

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_\Theta \theta p(\theta|y)d\theta = g(y)$$

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_\Theta \theta p(\theta|y)d\theta = g(y)$$

Introduction

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y)d\theta = g(y)$$

Introduction

- Suppose $Y \sim p(y|\theta)$ and $p(\theta)$ is the prior distribution over θ
- Our objective is to make statistical inference about θ
- The posterior distribution is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}$$

- Objective is to estimate the posterior mean:

$$E(\theta|y) = \int_\Theta \theta p(\theta|y)d\theta = g(y)$$

- So we will be concerned with evaluating integrals of the form

$$E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- $p()$ is the posterior probability density
- We can generate finitely many random samples $(\theta^1, ..., \theta^n)$ from $p(\theta|y)$
- Approximate the integral with

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i)$$

- So we will be concerned with evaluating integrals of the form

$$E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- $p()$ is the posterior probability density

- We can generate finitely many random samples $(\theta^1, ..., \theta^n)$ from $p(\theta|y)$

- Approximate the integral with

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i)$$

Monte Carlo Integration

- So we will be concerned with evaluating integrals of the form

$$E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- $p()$ is the posterior probability density

- We can generate finitely many random samples $(\theta^1, ..., \theta^n)$ from $p(\theta|y)$

- Approximate the integral with

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i)$$

- So we will be concerned with evaluating integrals of the form

$$E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- $p()$ is the posterior probability density
- We can generate finitely many random samples $(\theta^1, ..., \theta^n)$ from $p(\theta|y)$
- Approximate the integral with

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i)$$

- Convergence:

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i) \xrightarrow{a.s.} E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- CLT:

$$\frac{\bar{h} - E(\theta|y)}{\sqrt{\sigma_n^2}} \xrightarrow{L} N(0,1)$$

where $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}[\bar{h} - E(\theta|y)]^2$, such that $\int_{\Theta} h(\theta)^2 p(\theta|y)d\theta < \infty$

- Convergence:

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta^i) \xrightarrow{a.s.} E(h(\theta)|y) = \int_{\Theta} h(\theta)p(\theta|y)d\theta$$

- CLT:

$$\frac{\bar{h} - E(\theta|y)}{\sqrt{\sigma_n^2}} \xrightarrow{L} N(0,1)$$

where $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}[\bar{h} - E(\theta|y)]^2$, such that $\int_{\Theta} h(\theta)^2 p(\theta|y)d\theta < \infty$

- The advantage of CLT is we can evaluate the Monte Carlo error
- It assumes σ_n^2 is the proper estimate of the variance of $\bar{h}_n$
- If σ_n^2 does not converge, converges too slowly, a CLT may not apply. In such cases we will not be able to estimate the Monte Carlo error.

- The advantage of CLT is we can evaluate the Monte Carlo error
- It assumes σ_n^2 is the proper estimate of the variance of $\bar{h}_n$
- If σ_n^2 does not converge, converges too slowly, a CLT may not apply. In such cases we will not be able to estimate the Monte Carlo error.

- The advantage of CLT is we can evaluate the Monte Carlo error
- It assumes σ_n^2 is the proper estimate of the variance of $\bar{h}_n$
- If σ_n^2 does not converge, converges too slowly, a CLT may not apply. In such cases we will not be able to estimate the Monte Carlo error.

- Importance sampling is based on an alternative formulation of the SLLN

- For notational convinience we assume $p(\theta|y) = f(\theta)$

$$\begin{aligned} E_f(h(\theta)) &= \int_\Theta h(\theta) f(\theta) d\theta \\ &= \int_\Theta h(\theta) \frac{f(\theta)}{g(\theta)} g(\theta) d\theta \\ &= E_g\left[\frac{h(\theta) f(\theta)}{g(\theta)}\right] \end{aligned}$$

  - f is the target density
  - g is the cadidate density

- Importance sampling is based on an alternative formulation of the SLLN
- For notational convinience we assume $p(\theta|y) = f(\theta)$

$$
\begin{aligned}
E_f(h(\theta)) &= \int_{\Theta} h(\theta) f(\theta) d\theta \\
&= \int_{\Theta} h(\theta) \frac{f(\theta)}{g(\theta)} g(\theta) d\theta \\
&= E_g\left[\frac{h(\theta) f(\theta)}{g(\theta)}\right]
\end{aligned}
$$

- f is the target density
- g is the cadidate density

- Importance sampling is based on an alternative formulation of the SLLN
- For notational convinience we assume $p(\theta|y) = f(\theta)$

$$\begin{aligned} E_f(h(\theta)) &= \int_\Theta h(\theta) f(\theta) d\theta \\ &= \int_\Theta h(\theta) \frac{f(\theta)}{g(\theta)} g(\theta) d\theta \\ &= E_g\left[\frac{h(\theta) f(\theta)}{g(\theta)}\right] \end{aligned}$$

  - f is the target density
  - g is the cadidate density

- So in 'Importance Sampling' all you do is first you generate samples from candidate distribution g

- Suppose $(\theta^1, \theta^2, ..., \theta^N)$ are the random samples generated from g

- You estimates

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta^i)}{g(\theta^i)}h(\theta^i)$$

and by virtue of SLLN we have

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta^i)}{g(\theta^i)}h(\theta^i) \longrightarrow E_g\left[\frac{h(\theta)f(\theta)}{g(\theta)}\right] = E_f(h(\theta))$$

- So in 'Importance Sampling' all you do is first you generate samples from candidate distribution g

- Suppose $(\theta^1, \theta^2, ..., \theta^N)$ are the random samples generated from g

- You estimates

$$\bar{h} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\theta^i)}{g(\theta^i)} h(\theta^i)$$

and by virtue of SLLN we have

$$\bar{h} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\theta^i)}{g(\theta^i)} h(\theta^i) \longrightarrow E_g\left[\frac{h(\theta) f(\theta)}{g(\theta)}\right] = E_f(h(\theta))$$

- So in 'Importance Sampling' all you do is first you generate samples from candidate distribution g
- Suppose $(\theta^1, \theta^2, ..., \theta^N)$ are the random samples generated from g
- You estimates

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta^i)}{g(\theta^i)}h(\theta^i)$$

and by virtue of SLLN we have

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta^i)}{g(\theta^i)}h(\theta^i) \longrightarrow E_g\left[\frac{h(\theta)f(\theta)}{g(\theta)}\right] = E_f(h(\theta))$$

- So in 'Importance Sampling' all you do is first you generate samples from candidate distribution g
- Suppose $(\theta^1, \theta^2, ..., \theta^N)$ are the random samples generated from g
- You estimates
$$\bar{h} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\theta^i)}{g(\theta^i)} h(\theta^i)$$
and by virtue of SLLN we have
$$\bar{h} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\theta^i)}{g(\theta^i)} h(\theta^i) \longrightarrow E_g \left[\frac{h(\theta) f(\theta)}{g(\theta)} \right] = E_f(h(\theta))$$

- The logic underlying importance sampling lies in a simple rearrangement of terms in the target integral and multiplying by 1:

$$\int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = \int h(\theta)\omega(\theta)g(\theta)d\theta$$

  here $g()$ is another density whose support is same as $f()$.

- $\omega()$ is called importance function

- a good importance function will be large when the integrand is large and small otherwise.

- The logic underlying importance sampling lies in a simple rearrangement of terms in the target integral and multiplying by 1:

$$\int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = \int h(\theta)\omega(\theta)g(\theta)d\theta$$

  here $g()$ is another density whose support is same as $f()$.

- $\omega()$ is called importance function

- a good importance function will be large when the integrand is large and small otherwise.

Importance Sampling

- The logic underlying importance sampling lies in a simple rearrangement of terms in the target integral and multiplying by 1:

$$\int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = \int h(\theta)\omega(\theta)g(\theta)d\theta$$

  here $g()$ is another density whose support is same as $f()$.

- $\omega()$ is called importance function

- a good importance function will be large when the integrand is large and small otherwise.

- IS to improve integral approximation - it reduces the variance of an integral approximation.

- Another objective of IS when you cannot generate from the density p

- A third objective of IS is to draw inference or generate sample when the target density is unnormalized

Importance Sampling (IS)

- IS to improve integral approximation - it reduces the variance of an integral approximation.
- Another objective of IS when you cannot generate from the density p
- A third objective of IS is to draw inference or generate sample when the target density is unnormalized

- IS to improve integral approximation - it reduces the variance of an integral approximation.
- Another objective of IS when you cannot generate from the density p
- A third objective of IS is to draw inference or generate sample when the target density is unnormalized

Importance Sampling to improve integral approximation

Improve integral approximation

Consider the function $h(x) = 10\exp(-2|x-5|)$. Suppose that we want to calculate $E(h(X))$, where $X \sim Uniform(0, 10)$. That is we want to calculate

$$\int_0^{10} \exp(-2|x-5|)dx.$$

The true value for this integral is about 1.

- The simple way to do this is to generate x_i from the $Uniform(0, 10)$ and look at the sample mean of $h(x_i)$
- Notice this is equivalent to importance sampling with importance function $\omega(x) = f(x)$, where $f(x) = \frac{1}{10} I(0 < x < 10)$

Improve integral approximation

Consider the function $h(x) = 10\exp(-2|x-5|)$. Suppose that we want to calculate $E(h(X))$, where $X \sim Uniform(0,10)$. That is we want to calculate

$$\int_0^{10} \exp(-2|x-5|)dx.$$

The true value for this integral is about 1.

- The simple way to do this is to generate x_i from the $Uniform(0,10)$ and look at the sample mean of $h(x_i)$
- Notice this is equivalent to importance sampling with importance function $\omega(x) = f(x)$, where $f(x) = \frac{1}{10}I(0 < x < 10)$

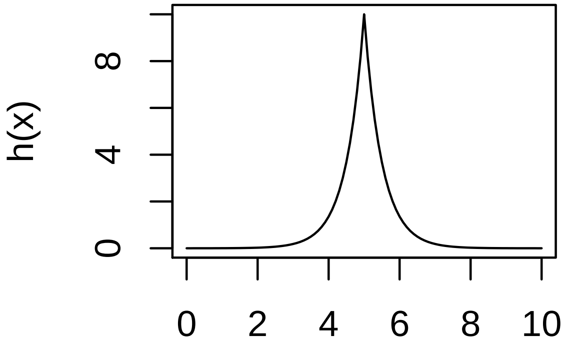
Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0,10)$

```
> set.seed(7831)
> sim.size<-50000
> X<-runif(sim.size,0,10)
> Y<-10*exp(-2*abs(X-5))
> c(mean(Y),var(Y))
[1] 0.999026 3.993018
```

Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0, 10)$



Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0, 10)$
- The function $h()$ in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation.
- Something more like a gaussian function ce^{-x^2} with a peak at 5 and small variance, say, 1, would provide greater precision
- We can re-write the integral as

$$\int_0^{10} 10\exp(-2|x-5|)\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0,10)$
- The function $h()$ in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation.
- Something more like a gaussian function ce^{-x^2} with a peak at 5 and small variance, say, 1, would provide greater precision
- We can re-write the integral as

$$\int_0^{10} 10\exp(-2|x-5|)\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0,10)$
- The function $h()$ in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation.
- Something more like a gaussian function ce^{-x^2} with a peak at 5 and small variance, say, 1, would provide greater precision
- We can re-write the integral as

$$\int_0^{10} 10\exp(-2|x-5|)\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

Importance Sampling to improve integral approximation

- Note that $h(x) = 10\exp(-2|x-5|)$ where $x \sim Unif(0,10)$
- The function $h()$ in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation.
- Something more like a gaussian function ce^{-x^2} with a peak at 5 and small variance, say, 1, would provide greater precision
- We can re-write the integral as

$$\int_0^{10} 10\exp(-2|x-5|)\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

Importance Sampling to improve integral approximation

- We want to estimate $E[h(X)\omega(X)]$ where $X \sim N(5,1)$
- So in this case $f(x) \sim Uniform(0,10)$, $g(x) \sim N(5,1)$
- The weight function is

$$\omega(x) = \frac{\sqrt{2\pi}e^{(x-5)^2/2}}{10}$$

Importance Sampling to improve integral approximation

- We want to estimate $E[h(X)\omega(X)]$ where $X \sim N(5,1)$
- So in this case $f(x) \sim Uniform(0,10)$, $g(x) \sim N(5,1)$
- The weight function is

$$\omega(x) = \frac{\sqrt{2\pi}e^{(x-5)^2/2}}{10}$$

Importance Sampling to improve integral approximation

- We want to estimate $E[h(X)\omega(X)]$ where $X \sim N(5, 1)$
- So in this case $f(x) \sim Uniform(0, 10)$, $g(x) \sim N(5, 1)$
- The weight function is

$$\omega(x) = \frac{\sqrt{2\pi}e^{(x-5)^2/2}}{10}$$

Importance Sampling to improve integral approximation

- We implement the IS approach as

```
> w<-function(x)dunif(x,0,10)/dnorm(x,mean=5,sd=1)
> h<-function(x)10*exp(-2*abs(x-5))
> set.seed(7831)
> sim.size<-50000
> x<-rnorm(sim.size,mean=5,sd=1)
> y<-h(x)*w(x)
> c(mean(y),var(y))
[1] 0.9995830 0.3609514
```

Note that here Monte Carlo variance is much smaller

Application:

Suppose daily return of a stock is defined as

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100$$

where P_t is the price of the stock on t^{th} day. Suppose it is believed that return of the stock follows $r_t \sim N(0, 1)$.

Objective: We want to estimate $P(r_t < -5)$, i.e., we want to estimate the probability that the price of the stock will drop more than 5% in one day?

Importance Sampling - Application

- As we are interested in $P(r_t < -5)$ where $r_t \sim N(0,1)$, the probability in log-scales, i.e., $\log(P(r_t < -5))$

```
> log(pnorm(-5,mean=0,sd=1))
[1] -15.065
```

- Simulating $Z^{(i)}$ from $N(0,1)$ only produces a hit once in about 3 million iterations !

- This is a very rare event for standard normal distribution

- Estimates from Importance sampling

```
[1] -15.0732
```

- As we are interested in $P(r_t < -5)$ where $r_t \sim N(0,1)$, the probability in log-scales, i.e., $\log(P(r_t < -5))$

```
> log(pnorm(-5,mean=0,sd=1))
[1] -15.065
```

- Simulating $Z^{(i)}$ from $N(0,1)$ only produces a hit once in about 3 million iterations !

- This is a very rare event for standard normal distribution

- Estimates from Importance sampling

```
[1] -15.0732
```

- As we are interested in $P(r_t < -5)$ where $r_t \sim N(0,1)$, the probability in log-scales, i.e., $\log(P(r_t < -5))$

  ```
  > log(pnorm(-5,mean=0,sd=1))
  [1] -15.065
  ```

- Simulating $Z^{(i)}$ from $N(0,1)$ only produces a hit once in about 3 million iterations !

- This is a very rare event for standard normal distribution

- Estimates from Importance sampling

  ```
  [1] -15.0732
  ```

- As we are interested in $P(r_t < -5)$ where $r_t \sim N(0,1)$, the probability in log-scales, i.e., $\log(P(r_t < -5))$

```
> log(pnorm(-5,mean=0,sd=1))
[1] -15.065
```

- Simulating $Z^{(i)}$ from $N(0,1)$ only produces a hit once in about 3 million iterations !
- This is a very rare event for standard normal distribution
- Estimates from Importance sampling

```
[1] -15.0732
```

Importance Sampling - Application

	Built-in R	Importance Sampling
$\log(P(r_t < -1))$	-1.841	-1.865
$\log(P(r_t < -2))$	-3.783	-3.803
$\log(P(r_t < -3))$	-6.608	-6.624
$\log(P(r_t < -4))$	-10.360	-10.372
$\log(P(r_t < -5))$	-15.065	-15.073

athshala पाठशाला
MHRD
Govt. of India
सत्यमेव जयते

- In Bayesian inference is the primary example where you want to determine the properties of a probability distribution given upto its kernel function.

- Suppose $k(\theta)$ is the kernel function of the posterior distribution $p(\theta|y)$, i.e.,

$$f(\theta) = p(\theta|y) = \frac{k(\theta)}{C},$$

where $C = \int k(\theta)d\theta = \int p(y|\theta)p(\theta)d\theta$

- Typically we do not know C and we know the posterior distribution in its unnormalized form, i.e.,

$$p(\theta|y) \propto p(y|\theta)p(\theta) = k(\theta)$$

- In Bayesian inference is the primary example where you want to determine the properties of a probability distribution given upto its kernel function.

- Suppose $k(\theta)$ is the kernel function of the posterior distribution $p(\theta|y)$, i.e.,

$$f(\theta) = p(\theta|y) = \frac{k(\theta)}{C},$$

  where $C = \int k(\theta)d\theta = \int p(y|\theta)p(\theta)d\theta$

- Typically we do not know C and we know the posterior distribution in its unnormalized form, i.e.,

$$p(\theta|y) \propto p(y|\theta)p(\theta) = k(\theta)$$

- In Bayesian inference is the primary example where you want to determine the properties of a probability distribution given upto its kernel function.

- Suppose $k(\theta)$ is the kernel function of the posterior distribution $p(\theta|y)$, i.e.,

$$f(\theta) = p(\theta|y) = \frac{k(\theta)}{C},$$

where $C = \int k(\theta)d\theta = \int p(y|\theta)p(\theta)d\theta$

- Typically we do not know C and we know the posterior distribution in its unnormalized form, i.e.,

$$p(\theta|y) \propto p(y|\theta)p(\theta) = k(\theta)$$

- If you want to calculate $E[h(\theta)]$ where unnormalized density of $f(\theta)$ is $k(\theta)$

- We can rewrite this as

$$\begin{aligned} E[h(\theta)] &= \int h(\theta)\frac{k(\theta)}{C}d(\theta) = \frac{1}{C}\int h(\theta)\frac{k(\theta)}{g(\theta)}g(\theta)d\theta \\ &= \frac{1}{C}\int h(\theta)\tilde{w}(\theta)g(\theta)d\theta \end{aligned}$$

where $\omega(\theta) = \frac{f(\theta)}{g(\theta)} = \frac{1}{C}\frac{k(\theta)}{g(\theta)} = \frac{1}{C}\tilde{\omega}(\theta)$ and $f(\theta) = p(\theta|y)$ is the normalized posterior density

- If you want to calculate $E[h(\theta)]$ where unnormalized density of $f(\theta)$ is $k(\theta)$

- We can rewrite this as

$$\begin{aligned} E[h(\theta)] &= \int h(\theta)\frac{k(\theta)}{C}d(\theta) = \frac{1}{C}\int h(\theta)\frac{k(\theta)}{g(\theta)}g(\theta)d\theta \\ &= \frac{1}{C}\int h(\theta)\tilde{w}(\theta)g(\theta)d\theta \end{aligned}$$

where $\omega(\theta) = \frac{f(\theta)}{g(\theta)} = \frac{1}{C}\frac{k(\theta)}{g(\theta)} = \frac{1}{C}\tilde{\omega}(\theta)$ and $f(\theta) = p(\theta|y)$ is the normalized posterior density

- By LLN if $\theta^1, \theta^2, ..., \theta^N$ are iid random samples from g, then

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i) \longrightarrow CE[h(\theta)]$$

as $N \longrightarrow \infty$.

- Also by LLN,

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}(\theta^i) \longrightarrow C$$

- Therefore Monte Carlo estimator for $E[h(\theta)]$ using Importance Sampling Scheme is

$$\bar{h}_{IS} = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i)}{\sum_{i=1}^{N} \tilde{\omega}(\theta^i)}$$

- By LLN if $\theta^1, \theta^2, ..., \theta^N$ are iid random samples from g, then

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i) \longrightarrow CE[h(\theta)]$$

as $N \longrightarrow \infty$.

- Also by LLN,

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}(\theta^i) \longrightarrow C$$

- Therefore Monte Carlo estimator for $E[h(\theta)]$ using Importance Sampling Scheme is

$$\bar{h}_{IS} = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i)}{\sum_{i=1}^{N} \tilde{\omega}(\theta^i)}$$

- By LLN if $\theta^1, \theta^2, ..., \theta^N$ are iid random samples from g, then

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i) \longrightarrow CE[h(\theta)]$$

as $N \longrightarrow \infty$.

- Also by LLN,

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}(\theta^i) \longrightarrow C$$

- Therefore Monte Carlo estimator for $E[h(\theta)]$ using Importance Sampling Scheme is

$$\bar{h}_{IS} = \frac{\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\tilde{\omega}(\theta^i)}{\sum_{i=1}^{N} \tilde{\omega}(\theta^i)}$$

- The method is only reliable when the weights are not too variable.

- As a rule of thumb, when

$$ESS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\omega}(\theta^i)}{\bar{\omega}} - 1\right)^2}$$

  is less than 5, the IS method is reasonable. Note that $\bar{\omega} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}(\theta^i)$

- You will know you chose a bad g when ESS is large.

Importance Sampling for Bayesian Inference

- The method is only reliable when the weights are not too variable.

- As a rule of thumb, when

$$ESS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\omega}(\theta^i)}{\bar{\omega}} - 1\right)^2}$$

is less than 5, the IS method is reasonable. Note that $\bar{\omega} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}(\theta^i)$

- You will know you chose a bad g when ESS is large.

- The method is only reliable when the weights are not too variable.

- As a rule of thumb, when

$$ESS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\omega}(\theta^i)}{\bar{\omega}} - 1\right)^2}$$

is less than 5, the IS method is reasonable. Note that $\bar{\omega} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}(\theta^i)$

- You will know you chose a bad g when ESS is large.

- The method is only reliable when the weights are not too variable.

- As a rule of thumb, when

$$ESS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\omega}(\theta^i)}{\bar{\omega}} - 1\right)^2}$$

is less than 5, the IS method is reasonable. Note that $\bar{\omega} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}(\theta^i)$

- You will know you chose a bad g when ESS is large.

- When $ESS < 5$ the variance $\bar{h}_{IS}$ can be estimated as

$$\sigma_{IS}^2 = \frac{1}{N}\sum_{i=1}^{N}(\zeta^i - \bar{h}_{IS})^2$$

where

$$\zeta^i = \frac{\tilde{\omega}(\theta^i)h(\theta^i)}{\bar{\omega}}$$

e-Pathshala पाठशाला MHRD Govt. of India

- Suppose following data presents number of accidents every month for last 12 months in a particular stretch on natitonal highway 34 (NH34)
- $6, 2, 2, 1, 2, 1, 1, 2, 3, 5, 2, 1$
- We define y_i as the number of accident on i^{th} month, i.e., $y_1 = 6$, $y_2 = 2$,... $y_{12} = 1$.
- We assume $y_i \overset{iid}{\sim}$ Poisson(θ) and $\theta \sim$ log-Normal$(\mu = 2, \sigma = 1)$

- Suppose following data presents number of accidents every month for last 12 months in a particular stretch on natitonal highway 34 (NH34)

- $6, 2, 2, 1, 2, 1, 1, 2, 3, 5, 2, 1$

- We define y_i as the number of accident on i^{th} month, i.e., $y_1 = 6$, $y_2 = 2$,... $y_{12} = 1$.

- We assume $y_i \overset{iid}{\sim}$ Poisson(θ) and $\theta \sim$ log-Normal$(\mu = 2, \sigma = 1)$

Application - Modeling Road Accident

- Suppose following data presents number of accidents every month for last 12 months in a particular stretch on natitonal highway 34 (NH34)
- $6, 2, 2, 1, 2, 1, 1, 2, 3, 5, 2, 1$
- We define y_i as the number of accident on i^{th} month, i.e., $y_1 = 6$, $y_2 = 2$,... $y_{12} = 1$.
- We assume $y_i \overset{iid}{\sim}$ Poisson(θ) and $\theta \sim$ log-Normal$(\mu = 2, \sigma = 1)$

Application - Modeling Road Accident

- Suppose following data presents number of accidents every month for last 12 months in a particular stretch on natitonal highway 34 (NH34)

- $6, 2, 2, 1, 2, 1, 1, 2, 3, 5, 2, 1$

- We define y_i as the number of accident on i^{th} month, i.e., $y_1 = 6$, $y_2 = 2$,... $y_{12} = 1$.

- We assume $y_i \stackrel{iid}{\sim}$ Poisson(θ) and $\theta \sim$ log-Normal$(\mu = 2, \sigma = 1)$

- In this problem since a non-conjugate prior is being selected the posterior distribution is known upto normalizing constant.
- Simulating random samples from the unnormalized posterior distribution is difficult
- So we simulate from a candidate distribution and using the importance sampling method we estimate the θ as

```
[1] 2.737056
```

- In this problem since a non-conjugate prior is being selected the posterior distribution is known upto normalizing constant.
- Simulating random samples from the unnormalized posterior distribution is difficult
- So we simulate from a candidate distribution and using the importance sampling method we estimate the θ as

```
[1] 2.737056
```

- In this problem since a non-conjugate prior is being selected the posterior distribution is known upto normalizing constant.
- Simulating random samples from the unnormalized posterior distribution is difficult
- So we simulate from a candidate distribution and using the importance sampling method we estimate the θ as

```
[1] 2.737056
```

- Note that simulated $(\theta^1, ..., \theta^N)$ is not from the target density ... it is from the candidate density g
- IS is an estimation method
- Later we will see that simulation from target can be done by adding step to IS

Pathshala पाठशाला MHRD Govt. of India

- Note that simulated $(\theta^1, ..., \theta^N)$ is not from the target density ... it is from the candidate density g
- IS is an estimation method
- Later we will see that simulation from target can be done by adding step to IS

- Note that simulated $(\theta^1, ..., \theta^N)$ is not from the target density ... it is from the candidate density g
- IS is an estimation method
- Later we will see that simulation from target can be done by adding step to IS

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Markov Chain Monte Carlo

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

- Markov Chain a stochastic process in which future states are independent of past states given the present state
- Monte Carlo simulation technique
- Up until now, we have done Monte Carlo simulation to find integrals instead of doing it analytically. This process is called **Monte Carlo Integration**.

- Markov Chain a stochastic process in which future states are independent of past states given the present state
- Monte Carlo simulation technique
- Up until now, we have done Monte Carlo simulation to find integrals instead of doing it analytically. This process is called **Monte Carlo Integration**.

- Markov Chain a stochastic process in which future states are independent of past states given the present state
- Monte Carlo simulation technique
- Up until now, we have done Monte Carlo simulation to find integrals instead of doing it analytically. This process is called **Monte Carlo Integration**.

Monte Carlo Integration

- Suppose we have a distribution $f(\theta)$ (perhaps a posterior) from which we want to draw samples.

- To derive it analytically, we need to take integrals:

$$I = \int_{\Theta} h(\theta) f(\theta) d\theta$$

where $h(\theta)$ is some function of θ ($h(\theta) = \theta$ for the mean and $h(\theta) = (\theta - E(\theta))^2$ for the variance).

- Suppose we have a distribution $f(\theta)$ (perhaps a posterior) from which we want to draw samples.

- To derive it analytically, we need to take integrals:

$$I = \int_{\Theta} h(\theta) f(\theta) d\theta$$

where $h(\theta)$ is some function of θ ($h(\theta) = \theta$ for the mean and $h(\theta) = (\theta - E(\theta))^2$ for the variance).

- We can approximate the integrals via Monte Carlo Integration by simulating M values from $p(\theta)$ and calculating

$$\hat{I} = \frac{1}{M} \sum_{i=1}^{M} h(\theta^i)$$

- For example, we can compute the expected value of the $Beta(3,4)$ distribution analytically:

$$E(\theta) = \int_0^1 \theta f(\theta) d\theta = \int_0^1 \theta \theta^{3-1}(1-\theta)^{4-1} d\theta = \frac{3}{7} = 0.4286$$

or via Monte Carlo methods:

```
> set.seed(88732)
> M <- 50000
> beta.sims <- rbeta(M, 3, 4)
> sum(beta.sims)/M
[1] 0.4288105
```

- Our Monte Carlo approximation $\hat{I}$ is a simulation estimator such that $\hat{I} \longrightarrow I$ as $M \longrightarrow \infty$

  This follows from SLLN.

Strong Law of Large Number (SLLN)

- Let X_1,X_2, ... be a sequence of **independent** and identically distributed random variables, each having a finite mean $\mu = E(X_i)$.

  Then with probability 1,

$$\frac{X_1 + X_2 + ...X_M}{M} \longrightarrow \mu \quad as \quad M \longrightarrow \infty$$

- In the example, each simulated random sample was independent and distributed from the same $Beta(3,4)$ distribution.

- But what if we cannot generate draws that are **independent**?

Strong Law of Large Number (SLLN)

- Let X_1,X_2, ... be a sequence of **independent** and identically distributed random variables, each having a finite mean $\mu = E(X_i)$.

  Then with probability 1,

$$\frac{X_1 + X_2 + ...X_M}{M} \longrightarrow \mu \quad as \quad M \longrightarrow \infty$$

- In the example, each simulated random sample was independent and distributed from the same $Beta(3, 4)$ distribution.

- But what if we cannot generate draws that are **independent**?

Strong Law of Large Number (SLLN)

- Let X_1,X_2, ... be a sequence of **independent** and identically distributed random variables, each having a finite mean $\mu = E(X_i)$.

  Then with probability 1,

$$\frac{X_1 + X_2 + ...X_M}{M} \longrightarrow \mu \quad as \quad M \longrightarrow \infty$$

- In the example, each simulated random sample was independent and distributed from the same $Beta(3, 4)$ distribution.

- But what if we cannot generate draws that are **independent**?

Fail to Draw Independent Random Samples

- Suppose we want to draw from our posterior distribution $p(\theta|y)$, but we cannot draw independent sample from it.
- For example, we often do not know the normalizing constant.
- We may be able to sample draws from $p(\theta|y)$ that are slightly dependent.
- If we can draw slightly dependent samples using a **Markov chain**, then can we solve the integration?

Fail to Draw Independent Random Samples

- Suppose we want to draw from our posterior distribution $p(\theta|y)$, but we cannot draw independent sample from it.
- For example, we often do not know the normalizing constant.
- We may be able to sample draws from $p(\theta|y)$ that are slightly dependent.
- If we can draw slightly dependent samples using a **Markov chain**, then can we solve the integration?

Fail to Draw Independent Random Samples

- Suppose we want to draw from our posterior distribution $p(\theta|y)$, but we cannot draw independent sample from it.
- For example, we often do not know the normalizing constant.
- We may be able to sample draws from $p(\theta|y)$ that are slightly dependent.
- If we can draw slightly dependent samples using a **Markov chain**, then can we solve the integration?

Fail to Draw Independent Random Samples

- Suppose we want to draw from our posterior distribution $p(\theta|y)$, but we cannot draw independent sample from it.
- For example, we often do not know the normalizing constant.
- We may be able to sample draws from $p(\theta|y)$ that are slightly dependent.
- If we can draw slightly dependent samples using a **Markov chain**, then can we solve the integration?

What is a Markov Chain?

- Definition: Markov Chain is a stochastic process in which future states are independent of past states given the present state

- Stochastic process: a consecutive set of random (not deterministic) quantities defined on some known state space Θ.

- Think of Θ as our parameter space

- consecutive implies a time component, indexed by t.

What is a Markov Chain?

- Definition: Markov Chain is a stochastic process in which future states are independent of past states given the present state
- Stochastic process: a consecutive set of random (not deterministic) quantities defined on some known state space Θ.
- Think of Θ as our parameter space
- consecutive implies a time component, indexed by t.

What is a Markov Chain?

- Definition: Markov Chain is a stochastic process in which future states are independent of past states given the present state
- Stochastic process: a consecutive set of random (not deterministic) quantities defined on some known state space Θ.
- Think of Θ as our parameter space
- consecutive implies a time component, indexed by t.

What is a Markov Chain?

- Definition: Markov Chain is a stochastic process in which future states are independent of past states given the present state
- Stochastic process: a consecutive set of random (not deterministic) quantities defined on some known state space Θ.
- Think of Θ as our parameter space
- consecutive implies a time component, indexed by t.

What is a Markov Chain?

- Consider a draw of $\theta^{(t)}$ to be a state at iteration t.
- The next draw $\theta^{(t+1)}$ is dependent only on the current draw $\theta^{(t)}$, and not on any past draws.
- Markov Property:

$$p(\theta^{(t+1)}|\theta^{(t)}, \theta^{(t-1)}, ..., \theta^{(1)}) = p(\theta^{(t+1)}|\theta^{(t)})$$

What is a Markov Chain?

- Consider a draw of $\theta^{(t)}$ to be a state at iteration t.
- The next draw $\theta^{(t+1)}$ is dependent only on the current draw $\theta^{(t)}$, and not on any past draws.
- Markov Property:

$$p(\theta^{(t+1)}|\theta^{(t)}, \theta^{(t-1)}, ..., \theta^{(1)}) = p(\theta^{(t+1)}|\theta^{(t)})$$

- Consider a draw of $\theta^{(t)}$ to be a state at iteration t.
- The next draw $\theta^{(t+1)}$ is dependent only on the current draw $\theta^{(t)}$, and not on any past draws.
- Markov Property:

$$p(\theta^{(t+1)}|\theta^{(t)}, \theta^{(t-1)}, ..., \theta^{(1)}) = p(\theta^{(t+1)}|\theta^{(t)})$$

What is Markov Chain?

- So our Markov chain is a bunch of draws of θ that are each dependent on the previous draw.
- The chain wanders around the parameter space, remembering only where it has been in the last time.
- What are the rules that govern how the chain jumps from one state to another at each time?
- The rule that govern the jump are known as **transition kernel**.

What is Markov Chain?

- So our Markov chain is a bunch of draws of θ that are each dependent on the previous draw.
- The chain wanders around the parameter space, remembering only where it has been in the last time.
- What are the rules that govern how the chain jumps from one state to another at each time?
- The rule that govern the jump are known as **transition kernel**.

What is Markov Chain?

- So our Markov chain is a bunch of draws of θ that are each dependent on the previous draw.
- The chain wanders around the parameter space, remembering only where it has been in the last time.
- What are the rules that govern how the chain jumps from one state to another at each time?
- The rule that govern the jump are known as **transition kernel**.

- So our Markov chain is a bunch of draws of θ that are each dependent on the previous draw.
- The chain wanders around the parameter space, remembering only where it has been in the last time.
- What are the rules that govern how the chain jumps from one state to another at each time?
- The rule that govern the jump are known as **transition kernel**.

Transition Kernel

- For discrete state space (k possible states): a $k \times k$ matrix of transition probabilities.

- **Example**: Suppose $k = 3$. The 3×3 transition matrix **P** would be

$p(\theta_1^{(t=1)} \vert \theta_1^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_1^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_1^{(t)})$
$p(\theta_1^{(t=1)} \vert \theta_2^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_2^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_2^{(t)})$
$p(\theta_1^{(t=1)} \vert \theta_3^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_3^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_3^{(t)})$

  where the subscripts index the 3 possible values that θ can take.

- The rows sum to one and define a conditional probability mass function, conditional on the current state.

- The columns are the marginal probabilities of being in a certain state in the next time.

- For discrete state space (k possible states): a $k \times k$ matrix of transition probabilities.

- **Example**: Suppose $k = 3$. The 3×3 transition matrix **P** would be

$p(\theta_1^{(t=1)} \vert \theta_1^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_1^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_1^{(t)})$
$p(\theta_1^{(t=1)} \vert \theta_2^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_2^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_2^{(t)})$
$p(\theta_1^{(t=1)} \vert \theta_3^{(t)})$	$p(\theta_2^{(t=1)} \vert \theta_3^{(t)})$	$p(\theta_3^{(t=1)} \vert \theta_3^{(t)})$

  where the subscripts index the 3 possible values that θ can take.

- The rows sum to one and define a conditional probability mass function, conditional on the current state.

- The columns are the marginal probabilities of being in a certain state in the next time.

- For discrete state space (k possible states): a $k \times k$ matrix of transition probabilities.

- **Example**: Suppose $k = 3$. The 3×3 transition matrix **P** would be

$p(\theta_1^{(t=1)} \mid \theta_1^{(t)})$	$p(\theta_2^{(t=1)} \mid \theta_1^{(t)})$	$p(\theta_3^{(t=1)} \mid \theta_1^{(t)})$
$p(\theta_1^{(t=1)} \mid \theta_2^{(t)})$	$p(\theta_2^{(t=1)} \mid \theta_2^{(t)})$	$p(\theta_3^{(t=1)} \mid \theta_2^{(t)})$
$p(\theta_1^{(t=1)} \mid \theta_3^{(t)})$	$p(\theta_2^{(t=1)} \mid \theta_3^{(t)})$	$p(\theta_3^{(t=1)} \mid \theta_3^{(t)})$

  where the subscripts index the 3 possible values that θ can take.

- The rows sum to one and define a conditional probability mass function, conditional on the current state.

- The columns are the marginal probabilities of being in a certain state in the next time.

- For discrete state space (k possible states): a $k \times k$ matrix of transition probabilities.

- **Example**: Suppose $k = 3$. The 3×3 transition matrix **P** would be

$p(\theta_1^{(t=1)} \| \theta_1^{(t)})$	$p(\theta_2^{(t=1)} \| \theta_1^{(t)})$	$p(\theta_3^{(t=1)} \| \theta_1^{(t)})$
$p(\theta_1^{(t=1)} \| \theta_2^{(t)})$	$p(\theta_2^{(t=1)} \| \theta_2^{(t)})$	$p(\theta_3^{(t=1)} \| \theta_2^{(t)})$
$p(\theta_1^{(t=1)} \| \theta_3^{(t)})$	$p(\theta_2^{(t=1)} \| \theta_3^{(t)})$	$p(\theta_3^{(t=1)} \| \theta_3^{(t)})$

  where the subscripts index the 3 possible values that θ can take.

- The rows sum to one and define a conditional probability mass function, conditional on the current state.

- The columns are the marginal probabilities of being in a certain state in the next time.

Discrete example:

1. Define a starting distribution $\Pi^{(0)}$ (a $1 \times k$ vector of probabilities that sum to one).

- At iteration 1, our distribution $\Pi^{(1)}$ (from which $\theta^{(1)}$ is drawn) is

$$\Pi^{(1)}_{(1\times k)} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}$$

- At iteration 2, our distribution $\Pi^{(2)}$ (from which $\theta^{(2)}$ is drawn) is

$$\Pi^{(2)}_{(1\times k)} = \Pi^{(1)}_{(1\times k)} \times \mathbf{P}$$

- At iteration t, our distribution $\Pi^{(t)}$ (from which $\theta^{(t)}$ is drawn) is

$$\Pi^{(t)}_{(1\times k)} = \Pi^{(t-1)}_{(1\times k)} \times \mathbf{P} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}^{t}$$

Discrete example:

1. Define a starting distribution $\Pi^{(0)}$ (a $1 \times k$ vector of probabilities that sum to one).

- At iteration 1, our distribution $\Pi^{(1)}$ (from which $\theta^{(1)}$ is drawn) is

$$\Pi^{(1)}_{(1\times k)} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}$$

- At iteration 2, our distribution $\Pi^{(2)}$ (from which $\theta^{(2)}$ is drawn) is

$$\Pi^{(2)}_{(1\times k)} = \Pi^{(1)}_{(1\times k)} \times \mathbf{P}$$

- At iteration t, our distribution $\Pi^{(t)}$ (from which $\theta^{(t)}$ is drawn) is

$$\Pi^{(t)}_{(1\times k)} = \Pi^{(t-1)}_{(1\times k)} \times \mathbf{P} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}^{t}$$

Discrete example:

1. Define a starting distribution $\Pi^{(0)}$ (a $1 \times k$ vector of probabilities that sum to one).

- At iteration 1, our distribution $\Pi^{(1)}$ (from which $\theta^{(1)}$ is drawn) is

$$\Pi^{(1)}_{(1\times k)} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}$$

- At iteration 2, our distribution $\Pi^{(2)}$ (from which $\theta^{(2)}$ is drawn) is

$$\Pi^{(2)}_{(1\times k)} = \Pi^{(1)}_{(1\times k)} \times \mathbf{P}$$

- At iteration t, our distribution $\Pi^{(t)}$ (from which $\theta^{(t)}$ is drawn) is

$$\Pi^{(t)}_{(1\times k)} = \Pi^{(t-1)}_{(1\times k)} \times \mathbf{P} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}^t$$

Discrete example:

1. Define a starting distribution $\Pi^{(0)}$ (a $1 \times k$ vector of probabilities that sum to one).

- At iteration 1, our distribution $\Pi^{(1)}$ (from which $\theta^{(1)}$ is drawn) is

$$\Pi^{(1)}_{(1\times k)} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}$$

- At iteration 2, our distribution $\Pi^{(2)}$ (from which $\theta^{(2)}$ is drawn) is

$$\Pi^{(2)}_{(1\times k)} = \Pi^{(1)}_{(1\times k)} \times \mathbf{P}$$

- At iteration t, our distribution $\Pi^{(t)}$ (from which $\theta^{(t)}$ is drawn) is

$$\Pi^{(t)}_{(1\times k)} = \Pi^{(t-1)}_{(1\times k)} \times \mathbf{P} = \Pi^{(0)}_{(1\times k)} \times \mathbf{P}^t$$

Stationarity of Markov Chain

- Define a stationary distribution π to be some distribution Π such that $\pi = \pi\mathbf{P}$
- Here we blur the distinction between the Markov chain and its transtion matrix $\mathbf{P}$.
- If posterior distribution is proper then the Markov Chain of the MCMC algorithm will typically converge to $p(\theta|y)$ regardless of our starting points.
- So if we can devise a Markov chain whose stationary distribution Π is our desired posterior distribution $p(\theta|y)$ then we can run this chain to get draws that approximately from $p(\theta|y)$ once the chain has converged.

- Define a stationary distribution π to be some distribution Π such that $\pi = \pi\mathbf{P}$
- Here we blur the distinction between the Markov chain and its transtion matrix $\mathbf{P}$.
- If posterior distribution is proper then the Markov Chain of the MCMC algorithm will typically converge to $p(\theta|y)$ regardless of our starting points.
- So if we can devise a Markov chain whose stationary distribution Π is our desired posterior distribution $p(\theta|y)$ then we can run this chain to get draws that approximately from $p(\theta|y)$ once the chain has converged.

Stationarity of Markov Chain

- Define a stationary distribution π to be some distribution Π such that $\pi = \pi\mathbf{P}$
- Here we blur the distinction between the Markov chain and its transtion matrix $\mathbf{P}$.
- If posterior distribution is proper then the Markov Chain of the MCMC algorithm will typically converge to $p(\theta|y)$ regardless of our starting points.
- So if we can devise a Markov chain whose stationary distribution Π is our desired posterior distribution $p(\theta|y)$ then we can run this chain to get draws that approximately from $p(\theta|y)$ once the chain has converged.

- Define a stationary distribution π to be some distribution Π such that $\pi = \pi \mathbf{P}$
- Here we blur the distinction between the Markov chain and its transtion matrix $\mathbf{P}$.
- If posterior distribution is proper then the Markov Chain of the MCMC algorithm will typically converge to $p(\theta|y)$ regardless of our starting points.
- So if we can devise a Markov chain whose stationary distribution Π is our desired posterior distribution $p(\theta|y)$ then we can run this chain to get draws that approximately from $p(\theta|y)$ once the chain has converged.

Fundamental Theorem of Markov Chain

If a Markov chain **P** is *irreducible*, *aperiodic* and *positive recurret* then it has an unique stationary distribution π.

- This is the unique (normalized such that the entries sum to 1) left eigenvector of **P** with eigenvalue 1.
- In addition,

$$P^t(x,y) \longrightarrow \pi(y) \text{ as } t \longrightarrow \infty \text{ for all } x, y \in \Omega.$$

- In light of the Fundamental theorem of MC, we shall refer to an irreducible, aperiodic Markov chain as ergodic.
- So as long as we can simulate an ergodic Markov chain, irrespective of the starting point the chain will converge to target stationary distribution.
- However, the time it takes for the chain to converge may vary; depending on the starting point.
- As a best practice, it is advisable to throw out a certain number of the first draws, known as the **burn-in**.

Burn-In

- In light of the Fundamental theorem of MC, we shall refer to an irreducible, aperiodic Markov chain as ergodic.
- So as long as we can simulate an ergodic Markov chain, irrespective of the starting point the chain will converge to target stationary distribution.
- However, the time it takes for the chain to converge may vary; depending on the starting point.
- As a best practice, it is advisable to throw out a certain number of the first draws, known as the **burn-in**.

Burn-In

- In light of the Fundamental theorem of MC, we shall refer to an irreducible, aperiodic Markov chain as ergodic.
- So as long as we can simulate an ergodic Markov chain, irrespective of the starting point the chain will converge to target stationary distribution.
- However, the time it takes for the chain to converge may vary; depending on the starting point.
- As a best practice, it is advisable to throw out a certain number of the first draws, known as the **burn-in**.

Burn-In

- In light of the Fundamental theorem of MC, we shall refer to an irreducible, aperiodic Markov chain as ergodic.
- So as long as we can simulate an ergodic Markov chain, irrespective of the starting point the chain will converge to target stationary distribution.
- However, the time it takes for the chain to converge may vary; depending on the starting point.
- As a best practice, it is advisable to throw out a certain number of the first draws, known as the **burn-in**.

- Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from $p(\theta|y)$, so it seems like we should be able to use Monte Carlo Integration methods to find quantities of interest.

- **One Problem**: MC draws are not independent, which we required for Monte Carlo Integration to work (required condition for SLLN to work).

- The answer is **Ergodic Theorem**

- Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from $p(\theta|y)$, so it seems like we should be able to use Monte Carlo Integration methods to find quantities of interest.

- **One Problem**: MC draws are not independent, which we required for Monte Carlo Integration to work (required condition for SLLN to work).

- The answer is **Ergodic Theorem**

- Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from $p(\theta|y)$, so it seems like we should be able to use Monte Carlo Integration methods to find quantities of interest.

- **One Problem**: MC draws are not independent, which we required for Monte Carlo Integration to work (required condition for SLLN to work).

- The answer is **Ergodic Theorem**

Ergodic Theorem

Let $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$ be T values from Markov chain that is aperiodic, irreducible and positive recurrent (that is the chain is ergodic) and $E[g(\theta)] < \infty$ Then with probability 1

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \longrightarrow \int_{\Theta} g(\theta)\pi(\theta)d\theta$$

as $T \longrightarrow \infty$ where π is the stationary distribution.

- This is the Markov chain analog to the LLN
- But what does it mean for a chain to be aperiodic, irreducible, and positive recurrent?

Ergodic Theorem

Let $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$ be T values from Markov chain that is aperiodic, irreducible and positive recurrent (that is the chain is ergodic) and $E[g(\theta)] < \infty$ Then with probability 1

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \longrightarrow \int_{\Theta} g(\theta)\pi(\theta)d\theta$$

as $T \longrightarrow \infty$ where π is the stationary distribution.

- This is the Markov chain analog to the LLN
- But what does it mean for a chain to be aperiodic, irreducible, and positive recurrent?

Ergodic Theorem

Let $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$ be T values from Markov chain that is aperiodic, irreducible and positive recurrent (that is the chain is ergodic) and $E[g(\theta)] < \infty$ Then with probability 1

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \longrightarrow \int_{\Theta} g(\theta)\pi(\theta)d\theta$$

as $T \longrightarrow \infty$ where π is the stationary distribution.

- This is the Markov chain analog to the LLN
- But what does it mean for a chain to be aperiodic, irreducible, and positive recurrent?

- A Markov chain P is aperiodic if for all $x, y \in \Omega$ we have $gcd\{t : P^t(x, y) > 0\} = 1$.

- If the only length of time for which the chain repeats some cycle of values is the trivial case with cycle length equal to one then the chain is aperiodic.

- Intuitively we can say repeatation is allowed but as long as the chain is not repeating itself in an identical cycle, then the chain is aperiodic

Aperiodicity

- A Markov chain P is aperiodic if for all $x, y \in \Omega$ we have $gcd\{t : P^t(x, y) > 0\} = 1$.
- If the only length of time for which the chain repeats some cycle of values is the trivial case with cycle length equal to one then the chain is aperiodic.
- Intuitively we can say repeatation is allowed but as long as the chain is not repeating itself in an identical cycle, then the chain is aperiodic

Aperiodicity

- A Markov chain P is aperiodic if for all $x, y \in \Omega$ we have $gcd\{t : P^t(x,y) > 0\} = 1$.
- If the only length of time for which the chain repeats some cycle of values is the trivial case with cycle length equal to one then the chain is aperiodic.
- Intuitively we can say repeatation is allowed but as long as the chain is not repeating itself in an identical cycle, then the chain is aperiodic

Irreducibility

- A Markov chain is *irreducible* if it is possible go from any state to any other state (not necessarily in one step).
- A Markov chain **P** is *irreducible* if for all x, y, there exists some t such that $P^t(x, y) > 0$.
- We can show if **P** is *irreducible*, then **P** is aperiodic $\iff there exists t such that \mathsf{P}^t(x, y) > 0$ for all $x, y \in \Omega$.

Irreducibility

- A Markov chain is *irreducible* if it is possible go from any state to any other state (not necessarily in one step).

- A Markov chain **P** is *irreducible* if for all x, y, there exists some t such that $P^t(x, y) > 0$.

- We can show if **P** is *irreducible*, then **P** is aperiodic $\iff thereexiststsuchthat \mathsf{P}^t(x, y) > 0$ for all $x, y \in \Omega$.

Irreducibility

- A Markov chain is *irreducible* if it is possible go from any state to any other state (not necessarily in one step).

- A Markov chain **P** is *irreducible* if for all x, y, there exists some t such that $P^t(x, y) > 0$.

- We can show if **P** is *irreducible*, then **P** is aperiodic $\iff there exists t such that \mathsf{P}^t(x, y) > 0$ for all $x, y \in \Omega$.

Positive Recurrence

- A Markov chain is *recurrent* if for any given state i , if the chain starts at i, it will eventually return to i with probability 1.

- A Markov chain is *positive recurrent* if the expected return time to state *i* is finite; otherwise it is null recurrent.

- So if our Markov chain is aperiodic, irreducible, and positive recurrent then it is ergodic and the ergodic theorem allows us to do Monte Carlo Integration by calculating $E[g(\theta)]$ from draws, ignoring the dependence between draws.

Positive Recurrence

- A Markov chain is *recurrent* if for any given state i , if the chain starts at i, it will eventually return to i with probability 1.

- A Markov chain is *positive recurrent* if the expected return time to state *i* is finite; otherwise it is null recurrent.

- So if our Markov chain is aperiodic, irreducible, and positive recurrent then it is ergodic and the ergodic theorem allows us to do Monte Carlo Integration by calculating $E[g(\theta)]$ from draws, ignoring the dependence between draws.

Positive Recurrence

- A Markov chain is *recurrent* if for any given state i , if the chain starts at i, it will eventually return to i with probability 1.

- A Markov chain is *positive recurrent* if the expected return time to state *i* is finite; otherwise it is null recurrent.

- So if our Markov chain is aperiodic, irreducible, and positive recurrent then it is ergodic and the ergodic theorem allows us to do Monte Carlo Integration by calculating $E[g(\theta)]$ from draws, ignoring the dependence between draws.

What is MCMC?

- MCMC is a class of Monte Carlo methods in which we can simulate dependent sample that are approximately from a posterior probability distribution.

- We then take the draws and calculate quantities of interest for the posterior distribution.

- In Bayesian statistics, there are generally two MCMC algorithms that we use: the Gibbs Sampler and the Metropolis-Hastings algorithm.

What is MCMC?

- MCMC is a class of Monte Carlo methods in which we can simulate dependent sample that are approximately from a posterior probability distribution.

- We then take the draws and calculate quantities of interest for the posterior distribution.

- In Bayesian statistics, there are generally two MCMC algorithms that we use: the Gibbs Sampler and the Metropolis-Hastings algorithm.

- MCMC is a class of Monte Carlo methods in which we can simulate dependent sample that are approximately from a posterior probability distribution.
- We then take the draws and calculate quantities of interest for the posterior distribution.
- In Bayesian statistics, there are generally two MCMC algorithms that we use: the Gibbs Sampler and the Metropolis-Hastings algorithm.

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

Application

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

Application

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

Application

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta\, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta\, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

Application

- We want to model the relationship as

$$r_t^F = \alpha + \beta \, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

Application

- We want to model the relationship as

$$r_t^F = \alpha + \beta \, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

- Prior:
  - $\alpha \sim N(0, \sigma_\alpha^2)$
  - $\beta \sim N(1, \sigma_\beta^2)$
  - $\sigma^{-2} \sim Gamma(c_0/2, d_0/2)$
- ```
  > library(MCMCpack)
  > posterior  <- MCMCregress(FTSE~DAX
  +                           , b0=c(0,1)
  +                           , B0 = 0.1
  +                           , sigma.mu = 5
  +                           , sigma.var = 25
  +                           , data=log_return
  +                           , verbose=0)
  ```
  ```

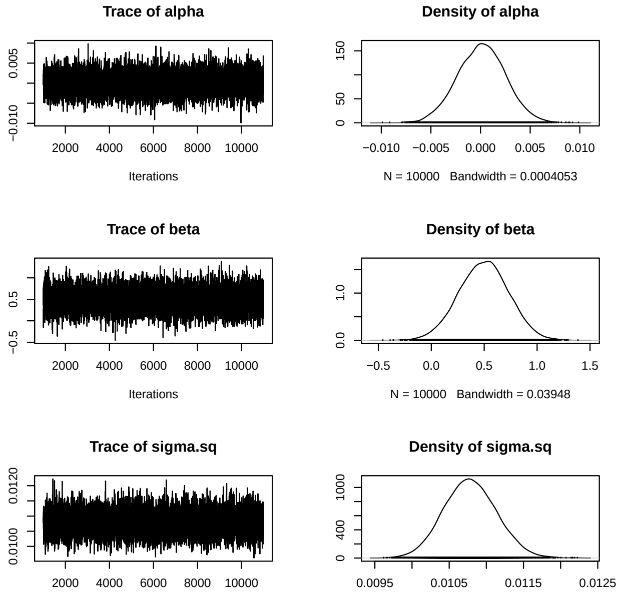
- Prior:
  - $\alpha \sim N(0, \sigma_\alpha^2)$
  - $\beta \sim N(1, \sigma_\beta^2)$
  - $\sigma^{-2} \sim Gamma(c_0/2, d_0/2)$
-
```
> library(MCMCpack)
> posterior  <- MCMCregress(FTSE~DAX
+                           , b0=c(0,1)
+                           , B0 = 0.1
+                           , sigma.mu = 5
+                           , sigma.var = 25
+                           , data=log_return
+                           , verbose=0)
```

Application

- Prior:
  - $\alpha \sim N(0, \sigma_\alpha^2)$
  - $\beta \sim N(1, \sigma_\beta^2)$
  - $\sigma^{-2} \sim Gamma(c_0/2, d_0/2)$
-
```
> library(MCMCpack)
> posterior  <- MCMCregress(FTSE~DAX
+                           , b0=c(0,1)
+                           , B0 = 0.1
+                           , sigma.mu = 5
+                           , sigma.var = 25
+                           , data=log_return
+                           , verbose=0)
```

- Prior:
  - $\alpha \sim N(0, \sigma_\alpha^2)$
  - $\beta \sim N(1, \sigma_\beta^2)$
  - $\sigma^{-2} \sim Gamma(c_0/2, d_0/2)$
-
```
> library(MCMCpack)
> posterior  <- MCMCregress(FTSE~DAX
+                                    , b0=c(0,1)
+                                    , B0 = 0.1
+                                    , sigma.mu = 5
+                                    , sigma.var = 25
+                                    , data=log_return
+                                    , verbose=0)
```

Application



Summary

	2.5%	25%	50%	75%	97.5%
alpha	-0.00465	-0.00152	0.00013	0.00174	0.00487
beta	0.03430	0.33824	0.49926	0.65432	0.94987
sigma.sq	0.01011	0.01054	0.01078	0.01102	0.01149

- 95% CI of α indicates that FTSE is fairly valued compare to DAX; since the interval include 0 in it.

- 95% CI of β indicates that the systematic risk of FTSE is less than DAX; since the interval doesnot include 1 or more precisely
  $P(\beta < 1|Data) = \frac{1}{T}\sum_{t=1}^{T} I(\beta^{(t)} < 1)$

```
> beta<-data.frame(posterior)$beta
> length(beta[beta<1])/length(beta)
[1] 0.985
```

e-Pathshala पाठशाला

A Gateway to all Postgraduate Courses

MHRD
Govt. of India
सत्यमेव जयते

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Regression Analysis - Part 1

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Introduction

- Bayesian Regression with conjugate and non-conjugate priors
- Set-up of the Bayesian Regression Model
- The standard “improper” non-informative prior
- Conjugate Prior Analysis

- Bayesian Regression with conjugate and non-conjugate priors
- Set-up of the Bayesian Regression Model
- The standard "improper" non-informative prior
- Conjugate Prior Analysis

- Bayesian Regression with conjugate and non-conjugate priors
- Set-up of the Bayesian Regression Model
- The standard “improper” non-informative prior
- Conjugate Prior Analysis

- Bayesian Regression with conjugate and non-conjugate priors
- Set-up of the Bayesian Regression Model
- The standard “improper” non-informative prior
- Conjugate Prior Analysis

- Many studies concern the relationship between two or more observable quantities.

- How do changes in quantity y (the dependent variable) vary as a function of another quantity x (the independent variable)?

- Regression models allow us to examine the conditional distribution of y given x, parameterized as $p(y|\beta, x)$ when the n observations (y_i, x_i) are exchangeable.

- Many studies concern the relationship between two or more observable quantities.
- How do changes in quantity y (the dependent variable) vary as a function of another quantity x (the independent variable)?
- Regression models allow us to examine the conditional distribution of y given x, parameterized as $p(y|\beta, x)$ when the n observations (y_i, x_i) are exchangeable.

Multiple Linear Regression

- Many studies concern the relationship between two or more observable quantities.
- How do changes in quantity y (the dependent variable) vary as a function of another quantity x (the independent variable)?
- Regression models allow us to examine the conditional distribution of y given x, parameterized as $p(y|\beta, x)$ when the n observations (y_i, x_i) are exchangeable.

- The normal linear model occurs when a distribution of y given x is normal with a mean equal to a linear function of X:

$$E(y_i|\beta, X) = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

for $i \in (1, 2, ..., n)$ and X_1 is a vector of one's.

- The ordinary linear regression model occurs when the variance of y given X, β is assumed to be constant over all observations.

- In other words, we have an ordinary linear regression model when:

$$y_i \sim N(\beta_1 + \beta_{2i} X_{2i} + ... + \beta_{ki} X_{ki}, \sigma^2)$$

for $i \in (1, 2, ..., n)$.

- The normal linear model occurs when a distribution of y given x is normal with a mean equal to a linear function of X:

$$E(y_i|\beta, X) = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

for $i \in (1, 2, ..., n)$ and X_1 is a vector of one's.

- The ordinary linear regression model occurs when the variance of y given X, β is assumed to be constant over all observations.

- In other words, we have an ordinary linear regression model when:

$$y_i \sim N(\beta_1 + \beta_{2i} X_{2i} + ... + \beta_{ki} X_{ki}, \sigma^2)$$

for $i \in (1, 2, ..., n)$.

- The normal linear model occurs when a distribution of y given x is normal with a mean equal to a linear function of X:

$$E(y_i|\beta, X) = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

for $i \in (1, 2, ..., n)$ and X_1 is a vector of one's.

- The ordinary linear regression model occurs when the variance of y given X, β is assumed to be constant over all observations.

- In other words, we have an ordinary linear regression model when:

$$y_i \sim N(\beta_1 + \beta_{2i} X_{2i} + ... + \beta_{ki} X_{ki}, \sigma^2)$$

for $i \in (1, 2, ..., n)$.

- If $y_i \sim N(\beta_1 + \beta_{2i}X_{2i} + ... + \beta_{ki}X_{ki}, \sigma^2)$ then it is well known that the ordinary least squares estimates and the maximum likelihood estimates of the parameters $\beta_1, ..., \beta_k$ are equivalent.

- If $\beta = [\beta_1, ..., \beta_k]^T$ then the frequentist estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- We know by the Gauss-Markov theorem that this estimate is BLUE.

- The frequentist estimate of σ^2 is $s^s = (y - X\beta)^T (y - X\beta)/(n - k)$

- If $y_i \sim N(\beta_1 + \beta_{2i}X_{2i} + ... + \beta_{ki}X_{ki}, \sigma^2)$ then it is well known that the ordinary least squares estimates and the maximum likelihood estimates of the parameters $\beta_1, ..., \beta_k$ are equivalent.

- If $\beta = [\beta_1, ..., \beta_k]^T$ then the frequentist estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- We know by the Gauss-Markov theorem that this estimate is BLUE.

- The frequentist estimate of σ^2 is $s^s = (y - X\beta)^T (y - X\beta)/(n - k)$

- If $y_i \sim N(\beta_1 + \beta_{2i}X_{2i} + ... + \beta_{ki}X_{ki}, \sigma^2)$ then it is well known that the ordinary least squares estimates and the maximum likelihood estimates of the parameters $\beta_1, ..., \beta_k$ are equivalent.
- If $\beta = [\beta_1, ..., \beta_k]^T$ then the frequentist estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- We know by the Gauss-Markov theorem that this estimate is BLUE.
- The frequentist estimate of σ^2 is $s^s = (y - X\beta)^T (y - X\beta)/(n - k)$

- If $y_i \sim N(\beta_1 + \beta_{2i}X_{2i} + ... + \beta_{ki}X_{ki}, \sigma^2)$ then it is well known that the ordinary least squares estimates and the maximum likelihood estimates of the parameters $\beta_1, ..., \beta_k$ are equivalent.

- If $\beta = [\beta_1, ..., \beta_k]^T$ then the frequentist estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- We know by the Gauss-Markov theorem that this estimate is BLUE.

- The frequentist estimate of σ^2 is
$s^s = (y - X\beta)^T (y - X\beta)/(n - k)$

- The uncertainty about the quantities β is summarized by the regression coefficients standard error which is the diagonal of the matrix: $Var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$

- Finally, we know that if V_i is the i^{th} diagonal element of $Var(\hat{\beta})$, then : $(\hat{\beta} - 0)/\sqrt{sV_i} \sim t_{n-k}$

- This statistic forms the basis for our hypothesis tests.

- The uncertainty about the quantities β is summarized by the regression coefficients standard error which is the diagonal of the matrix: $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
- Finally, we know that if V_i is the i^{th} diagonal element of $Var(\hat{\beta})$, then : $(\hat{\beta} - 0)/\sqrt{sV_i} \sim t_{n-k}$
- This statistic forms the basis for our hypothesis tests.

- The uncertainty about the quantities β is summarized by the regression coefficients standard error which is the diagonal of the matrix: $Var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$
- Finally, we know that if V_i is the i^{th} diagonal element of $Var(\hat{\beta})$, then : $(\hat{\beta} - 0)/\sqrt{sV_i} \sim t_{n-k}$
- This statistic forms the basis for our hypothesis tests.

Bayesian Setup

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

Bayesian Setup

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- By Bayes Rule, the posterior distribution is:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- To make inferences about the regression coefficients, we obviously need to choose a prior distribution for β, σ^2

- The standard non-informative prior distribution is uniform on $(\beta, \log \sigma^2)$ which is equivalent to :

$$p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$$

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.

Bayesian Regression with Noninformative P

- By Bayes Rule, the posterior distribution is:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- To make inferences about the regression coefficients, we obviously need to choose a prior distribution for β, σ^2

- The standard non-informative prior distribution is uniform on $(\beta, \log \sigma^2)$ which is equivalent to :

$$p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$$

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.

- By Bayes Rule, the posterior distribution is:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- To make inferences about the regression coefficients, we obviously need to choose a prior distribution for β, σ^2

- The standard non-informative prior distribution is uniform on $(\beta, \log \sigma^2)$ which is equivalent to :

$$p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$$

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.

- By Bayes Rule, the posterior distribution is:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- To make inferences about the regression coefficients, we obviously need to choose a prior distribution for β, σ^2
- The standard non-informative prior distribution is uniform on $(\beta, \log \sigma^2)$ which is equivalent to :

$$p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$$

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.

Bayesian Regression with Noninformative P

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.
- Why?
- Because if you have a large lot of data and few parameters, then the likelihood function is very sharply peaked, which means that the likelihood (the data) will dominate posterior inferences.
- With small sample sizes or a lot of parameters, prior distributions or hierarchical models become more important for analysis.

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.
- Why?
- Because if you have a large lot of data and few parameters, then the likelihood function is very sharply peaked, which means that the likelihood (the data) will dominate posterior inferences.
- With small sample sizes or a lot of parameters, prior distributions or hierarchical models become more important for analysis.

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.
- Why?
- Because if you have a large lot of data and few parameters, then the likelihood function is very sharply peaked, which means that the likelihood (the data) will dominate posterior inferences.
- With small sample sizes or a lot of parameters, prior distributions or hierarchical models become more important for analysis.

- This prior is a good choice for statistical models when you have a lot of data points and only a few parameters.

- Why?

- Because if you have a large lot of data and few parameters, then the likelihood function is very sharply peaked, which means that the likelihood (the data) will dominate posterior inferences.

- With small sample sizes or a lot of parameters, prior distributions or hierarchical models become more important for analysis.

e-Pathshala पाठशाला MHRD Govt. of India सत्यमेव जयते

- If $y \sim N(X\beta, \sigma^2)$ and $p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$, then conditional posterior distribution is

$$p(\beta|\sigma^2, y, X) \sim MN_k(\hat{\beta}, \sigma^2(X^T X)^{-1})$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$

- This follows from by completing the square, like we have seen over and over again when dealing with the normal distribution.

- The posterior distribution of σ^2 can be written as:

$$p(\sigma^2|y, X) \sim Scaled - Inv\chi^2(n-k, s^2)$$

where $s^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})/(n-k)$

- If $y \sim N(X\beta, \sigma^2)$ and $p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$, then conditional posterior distribution is

$$p(\beta|\sigma^2, y, X) \sim MN_k(\hat{\beta}, \sigma^2(X^T X)^{-1})$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$

- This follows from by completing the square, like we have seen over and over again when dealing with the normal distribution.

- The posterior distribution of σ^2 can be written as:

$$p(\sigma^2|y, X) \sim Scaled - Inv\chi^2(n-k, s^2)$$

where $s^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/(n-k)$

- If $y \sim N(X\beta, \sigma^2)$ and $p(\beta, \log(\sigma^2)) \propto \sigma^{-2}$, then conditional posterior distribution is

$$p(\beta|\sigma^2, y, X) \sim MN_k(\hat{\beta}, \sigma^2(X^TX)^{-1})$$

where $\hat{\beta} = (X^TX)^{-1}X^Ty$

- This follows from by completing the square, like we have seen over and over again when dealing with the normal distribution.

- The posterior distribution of σ^2 can be written as:

$$p(\sigma^2|y, X) \sim Scaled - Inv\chi^2(n-k, s^2)$$

where $s^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/(n-k)$

- Integrating out σ^2 from the full posterior distribution the marginal posteror distribution of β follow Multivariate t-distribution, i.e.,

$$p(\beta|y, X) \sim t_{n-k}(\hat{\beta}, s^2(X^T X)^{-1})$$

- Notice the close comparison with the classical results. The key difference would be interpretation of the standard errors.

- Integrating out σ^2 from the full posterior distribution the marginal posteror distribution of β follow Multivariate t-distribution, i.e.,

$$p(\beta|y, X) \sim t_{n-k}(\hat{\beta}, s^2(X^T X)^{-1})$$

- Notice the close comparison with the classical results. The key difference would be interpretation of the standard errors.

e-Pathshala पाठशाला MHRD Govt. of India सत्यमेव जयते

- Model evaluation proceeds just like it would in the traditional OLS framework. Models are largely evaluated based on their fitted (predicted) values.

- For example,

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

where $\hat{y}_i$ is the fitted value of y.

- In this respect, Bayesian model evaluation is straight forward.

- The fitted value for an observation i is a random draw from a t distribution with mean $X_i\hat{\beta}$ and variance $s^2(I + X_i(X^TX)X_i^T)$ on $(n-k)$ degrees of freedom.

Bayesian Model Evaluation

- Model evaluation proceeds just like it would in the traditional OLS framework. Models are largely evaluated based on their fitted (predicted) values.

- For example,

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

where $\hat{y}_i$ is the fitted value of y.

- In this respect, Bayesian model evaluation is straight forward.

- The fitted value for an observation i is a random draw from a t distribution with mean $X_i\hat{\beta}$ and variance $s^2(I + X_i(X^TX)X_i^T)$ on $(n-k)$ degrees of freedom.

Bayesian Model Evaluation

- Model evaluation proceeds just like it would in the traditional OLS framework. Models are largely evaluated based on their fitted (predicted) values.

- For example,

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

where $\hat{y}_i$ is the fitted value of y.

- In this respect, Bayesian model evaluation is straight forward.

- The fitted value for an observation i is a random draw from a t distribution with mean $X_i\hat{\beta}$ and variance $s^2(I + X_i(X^TX)X_i^T)$ on $(n-k)$ degrees of freedom.

Bayesian Model Evaluation

- Model evaluation proceeds just like it would in the traditional OLS framework. Models are largely evaluated based on their fitted (predicted) values.

- For example,
$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$
where $\hat{y}_i$ is the fitted value of y.

- In this respect, Bayesian model evaluation is straight forward.

- The fitted value for an observation i is a random draw from a t distribution with mean $X_i\hat{\beta}$ and variance $s^2(I + X_i(X^TX)X_i^T)$ on $(n-k)$ degrees of freedom.

- In most cases, you can simply use $X_i \hat{\beta}$ as the fitted value and ignore this additional uncertainty.

- Other diagnostics that you should use are plots of errors versus covariates histograms for the error term, etc. to make sure that the assumptions of the normal linear model hold.

- In most cases, you can simply use $X_i\hat{\beta}$ as the fitted value and ignore this additional uncertainty.

- Other diagnostics that you should use are plots of errors versus covariates histograms for the error term, etc. to make sure that the assumptions of the normal linear model hold.

Comment on the regression setup

There are two subtle points regarding the Bayesian regression setup.

- First, a full Bayesian model includes a distribution for the independent variable X, $p(X|\Psi)$.

  Therefore we have a joint likelihood $p(X, y|\Psi, \beta, \sigma)$ and joint prior $p(\Psi, \beta, \sigma)$.

  The fundamental assumption of the normal linear model is that $p(y|X, \beta, \sigma)$ and $p(X|\Psi)$ are independent in their prior distributions such that the posterior distribution factors into: $p(\Psi, \beta, \sigma|X, y) = p(\beta, \sigma|X, y)p(\Psi|X, y)$.

  As a result, $p(\beta, \sigma|X, y) \propto p(\beta, \sigma)p(y|\beta, \sigma, X)$

There are two subtle points regarding the Bayesian regression setup.

- First, a full Bayesian model includes a distribution for the independent variable X, $p(X|\Psi)$.

  Therefore we have a joint likelihood $p(X, y|\Psi, \beta, \sigma)$ and joint prior $p(\Psi, \beta, \sigma)$.

  The fundamental assumption of the normal linear model is that $p(y|X, \beta, \sigma)$ and $p(X|\Psi)$ are independent in their prior distributions such that the posterior distribution factors into: $p(\Psi, \beta, \sigma|X, y) = p(\beta, \sigma|X, y)p(\Psi|X, y)$.

  As a result, $p(\beta, \sigma|X, y) \propto p(\beta, \sigma)p(y|\beta, \sigma, X)$

There are two subtle points regarding the Bayesian regression setup.

- First, a full Bayesian model includes a distribution for the independent variable X, $p(X|\Psi)$.

  Therefore we have a joint likelihood $p(X, y|\Psi, \beta, \sigma)$ and joint prior $p(\Psi, \beta, \sigma)$.

  The fundamental assumption of the normal linear model is that $p(y|X, \beta, \sigma)$ and $p(X|\Psi)$ are independent in their prior distributions such that the posterior distribution factors into: $p(\Psi, \beta, \sigma|X, y) = p(\beta, \sigma|X, y)p(\Psi|X, y)$.

  As a result, $p(\beta, \sigma|X, y) \propto p(\beta, \sigma)p(y|\beta, \sigma, X)$

There are two subtle points regarding the Bayesian regression setup.

- Second, when we setup our probability model, we are implicitly conditioning on a model, call it H, which represents our beliefs about the data-generating process. Thus,

$$p(\beta, \sigma | X, y, H) \propto p(\beta, \sigma | H) p(y | \beta, \sigma, X, H)$$

- It is important to keep in mind that our inferences are dependent on H, and this is equally true for the frequentist perspective, where results can be dependent on the choice of likelihood function, covariates, etc.

Comment on the regression setup

There are two subtle points regarding the Bayesian regression setup.

- Second, when we setup our probability model, we are implicitly conditioning on a model, call it H, which represents our beliefs about the data-generating process. Thus,

$$p(\beta, \sigma | X, y, H) \propto p(\beta, \sigma | H) p(y | \beta, \sigma, X, H)$$

- It is important to keep in mind that our inferences are dependent on H, and this is equally true for the frequentist perspective, where results can be dependent on the choice of likelihood function, covariates, etc.

- Suppose that instead of an improper prior, we decide to use the conjugate prior.

- For the normal regression model, the conjugate prior distribution for $p(\beta_0, ..., \beta_k, \sigma^2)$ is the normal-inverse-gamma distribution.

- We have seen this distribution before when we studied the normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

- Suppose that instead of an improper prior, we decide to use the conjugate prior.
- For the normal regression model, the conjugate prior distribution for $p(\beta_0, ..., \beta_k, \sigma^2)$ is the normal-inverse-gamma distribution.
- We have seen this distribution before when we studied the normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

- Suppose that instead of an improper prior, we decide to use the conjugate prior.

- For the normal regression model, the conjugate prior distribution for $p(\beta_0, ..., \beta_k, \sigma^2)$ is the normal-inverse-gamma distribution.

- We have seen this distribution before when we studied the normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

  where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
  and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- Posterior mean:

$$E(\beta|y, X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y,X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y, X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y, X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Note: Zellner (1971) treats β_0 and conditional prior variance Λ_0 in the following way:

  suppose you have two data sets- (y_1, X_1) and (y_2, X_2). He sets β_0 equal to the posterior mean for a regression analysis of (y_1, X_1) with the improper prior $1/\sigma^2$ and set Λ_0 equal to $X_1^T X_1$.

- To summarize our uncertainty about the coefficients, the variance-covariance matrix for β is given by:

$$Cov(\beta|y, X) = \frac{\tilde{s}(\Lambda_0^{-1} + X^T X)^{-1}}{n + a - k - 3}$$

where

$$\begin{aligned} \tilde{s} = 2b \quad &+ \quad \frac{(y - X\beta)^T (y - X\beta)}{n + a - k - 3} \\ &+ \quad (\beta_0 - E(\beta))^T \Lambda_0^{-1} \beta_0 + (\hat{\beta} - E(\beta))^T X^T X \hat{\beta} \end{aligned}$$

- The posterior standard deviation can be taken from the square root of the diagonal of this matrix.

- The second term is the maximum likelihood estimate of the variance.
- The third terms states that our variance estimates will be greater if our prior values for the regression coefficients differ from their posterior values, especially if we indicate a great deal of confidence in our prior beliefs by assigning small variances in the matrix Λ_0.
- The fourth term states that our variance estimates for the regression coefficients will be greater if the standard OLS estimates differ from the posterior values, especially if $X^T X$ is a large number.

- The second term is the maximum likelihood estimate of the variance.
- The third terms states that our variance estimates will be greater if our prior values for the regression coefficients differ from their posterior values, especially if we indicate a great deal of confidence in our prior beliefs by assigning small variances in the matrix Λ_0.
- The fourth term states that our variance estimates for the regression coefficients will be greater if the standard OLS estimates differ from the posterior values, especially if $X^T X$ is a large number.

- The second term is the maximum likelihood estimate of the variance.

- The third terms states that our variance estimates will be greater if our prior values for the regression coefficients differ from their posterior values, especially if we indicate a great deal of confidence in our prior beliefs by assigning small variances in the matrix Λ_0.

- The fourth term states that our variance estimates for the regression coefficients will be greater if the standard OLS estimates differ from the posterior values, especially if $X^T X$ is a large number.

Application

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

Application

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

- DAX and FTSE are two stock market indexes for German and UK stock exchanges respectively.
- Suppose P_t^D is the DAX value on t^{th}-day
- and P_t^F is the FTSE value on t^{th}-day
- Corresponding log-return is
- $r_t^D = \log(P_t^D) - \log(P_{t-1}^D)$
- $r_t^F = \log(P_t^F) - \log(P_{t-1}^F)$

- We want to model the relationship as

$$r_t^F = \alpha + \beta\ r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

Application

- We want to model the relationship as

$$r_t^F = \alpha + \beta\, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

- We want to model the relationship as

$$r_t^F = \alpha + \beta\, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\alpha > 0$ means FTSE is undervalued compare to DAX
- $\alpha < 0$ means FTSE is overvalued compare to DAX
- $\alpha = 0$ means FTSE is fairly valued compare to DAX

Application

- We want to model the relationship as

$$r_t^F = \alpha + \beta \; r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

We consider only first 20 days of return, $n = 20$

- We want to model the relationship as

$$r_t^F = \alpha + \beta \, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX

- $\beta < 1$ means systematic risk of FTSE is less than DAX

- $\beta = 1$ means systematic risk of FTSE is equal to DAX

We consider only first 20 days of return, $n = 20$

- We want to model the relationship as

$$r_t^F = \alpha + \beta \, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

- $\beta > 1$ means systematic risk of FTSE is more than DAX
- $\beta < 1$ means systematic risk of FTSE is less than DAX
- $\beta = 1$ means systematic risk of FTSE is equal to DAX

We consider only first 20 days of return, $n = 20$

- We want to model the relationship as

$$r_t^F = \alpha + \beta \, r_t^D + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$

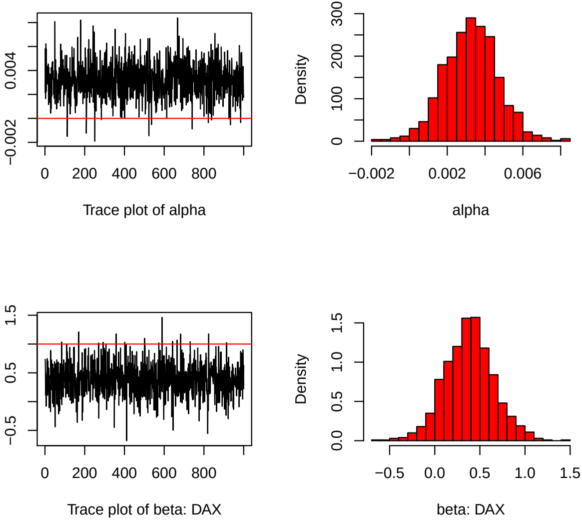
- $\beta > 1$ means systematic risk of FTSE is more than DAX

- $\beta < 1$ means systematic risk of FTSE is less than DAX

- $\beta = 1$ means systematic risk of FTSE is equal to DAX

We consider only first 20 days of return, $n = 20$

Application



Summary

```
       alpha  beta:DAX
2.5%  0.00040 -0.15973
50%   0.00327  0.38210
97.5% 0.00600  0.94316
```

- 95% CI of α indicates that FTSE is under valued compare to DAX; since the interval is entirely above 0.
- 95% CI of β indicates that the systematic risk of FTSE is less than DAX; since the interval doesnot include 1 or more precisely

  $P(\beta < 1|Data) = \frac{1}{T}\sum_{t=1}^{T} I(\beta^{(t)} < 1)$

```
> length(b.star[b.star[,2]<1,2])/1000
[1] 0.984
```

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Regression Analysis - Part 2

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Bayesian Setup

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

Bayesian Setup

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

Bayesian Setup

- For the normal linear model, we have:

$$y_i \sim N(\mu_i, \sigma^2)$$

for $i \in (1, 2, 3, ..., n)$ where μ_i is just an indicator for the expression:

$$\mu_i = \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- The object of statistical inference is the posterior distribution of the parameters $\beta_1 ..., \beta_k$ and σ^2

- By Bayes Rule, we know that this is simply:

$$p(\beta_1, ..., \beta_k, \sigma^2 | y, X) \propto \prod_{i=1}^{n} p(y_i | \mu_i, \sigma^2) p(\beta_1, ..., \beta_k, \sigma^2)$$

- For the normal regression model, the conjugate prior distribution for $p(\beta_0, ..., \beta_k, \sigma^2)$ is the normal-inverse-gamma distribution.

- We have seen this distribution before when we studied the normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

- For the normal regression model, the conjugate prior distribution for $p(\beta_0, ..., \beta_k, \sigma^2)$ is the normal-inverse-gamma distribution.

- We have seen this distribution before when we studied the normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- As we have studied normal model with unknown mean and variance. We know that this distribution can be factored such that:

$$p(\beta_0, ..., \beta_k, \sigma^2) = p(\beta_0, ..., \beta_k|\sigma^2)p(\sigma^2)$$

where $p(\beta_0, ..., \beta_k|\sigma^2) \sim MN_k(\beta_0, \Lambda_0)$
and $p(\sigma^2) \sim Inv - Gamma(a_0, b_0)$

- If we use a conjugate prior, then the prior distribution will have the same form. Thus, the posterior distribution will also follow normal-inverse-gamma.

- If we integrate out σ^2 the marginal for β will be a multivariate t-distribution.

- Posterior mean:

$$E(\beta|y,X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y,X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y,X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Posterior mean:

$$E(\beta|y, X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

- Notice that the coefficients are essentially a weighted average of the prior coefficients described by β_0 and standard OLS estimate $\hat{\beta}$.

- The weights are provided by the conditional prior precision Λ_0^{-1} and the data $X^T X$.

- This should make clear that as we increase our prior precision (decrease our prior variance) for β we place greater posterior weight on our prior beliefs relative to the data.

- Note: Zellner (1971) treats β_0 and conditional prior variance Λ_0 in the following way:

  suppose you have two data sets- (y_1, X_1) and (y_2, X_2). He sets β_0 equal to the posterior mean for a regression analysis of (y_1, X_1) with the improper prior $1/\sigma^2$ and set Λ_0 equal to $X_1^T X_1$.

- To summarize our uncertainty about the coefficients, the variance-covariance matrix for β is given by:

$$Cov(\beta|y,X) = \frac{\tilde{s}(\Lambda_0^{-1} + X^T X)^{-1}}{n+a-k-3}$$

where

$$\begin{aligned} \tilde{s} = 2b \quad &+ \quad \frac{(y - X\beta)^T (y - X\beta)}{n+a-k-3} \\ &+ \quad (\beta_0 - E(\beta))^T \Lambda_0^{-1} \beta_0 + (\hat{\beta} - E(\beta))^T X^T X \hat{\beta} \end{aligned}$$

- The posterior standard deviation can be taken from the square root of the diagonal of this matrix.

Is there a fairer test?

- With the small sample size, it is difficult to say conclusively that the parties are not following their activists.
- We can use the Bayesian method and incorporate prior probability about the data-generating process to the model.

Is there a fairer test?

- With the small sample size, it is difficult to say conclusively that the parties are not following their activists.
- We can use the Bayesian method and incorporate prior probability about the data-generating process to the model.

Justification for the Bayesian Approach

- Wallerstein and Stephens reach an empirical impass, where they do not able to adjudicate between the two theories, because of the small sample size and multicollinear variables.

- The incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients.

- Priors can be developed as equivalent to prior data sets, inflating the de facto n.

- The data set contains all available observations from a population of interest-it is not a random sample. More generally, cross-national data sets are not generated by a "repeatable" data-generating process.

- Wallerstein and Stephens reach an empirical impass, where they do not able to adjudicate between the two theories, because of the small sample size and multicollinear variables.

- The incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients.

- Priors can be developed as equivalent to prior data sets, inflating the de facto n.

- The data set contains all available observations from a population of interest-it is not a random sample. More generally, cross-national data sets are not generated by a "repeatable" data-generating process.

Justification for the Bayesian Approach

- Wallerstein and Stephens reach an empirical impass, where they do not able to adjudicate between the two theories, because of the small sample size and multicollinear variables.

- The incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients.

- Priors can be developed as equivalent to prior data sets, inflating the de facto n.

- The data set contains all available observations from a population of interest-it is not a random sample. More generally, cross-national data sets are not generated by a "repeatable" data-generating process.

Justification for the Bayesian Approach

- Wallerstein and Stephens reach an empirical impass, where they do not able to adjudicate between the two theories, because of the small sample size and multicollinear variables.

- The incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients.

- Priors can be developed as equivalent to prior data sets, inflating the de facto n.

- The data set contains all available observations from a population of interest-it is not a random sample. More generally, cross-national data sets are not generated by a "repeatable" data-generating process.

Justification for the Bayesian Approach

- Frequentist inference about a statistic (e.g. a regression coef.) is obtained through the assumption that the process generating the data could be repeated a large number of times.

- Specifically, frequentist inference is about the proportion of the time that, in the long-run, realizations of this statistic will fall within some interval.

- If there is no long-run, or possibility of repetition, then probabilistic summaries are not appropriate.

Justification for the Bayesian Approach

- Frequentist inference about a statistic (e.g. a regression coef.) is obtained through the assumption that the process generating the data could be repeated a large number of times.
- Specifically, frequentist inference is about the proportion of the time that, in the long-run, realizations of this statistic will fall within some interval.
- If there is no long-run, or possibility of repetition, then probabilistic summaries are not appropriate.

Justification for the Bayesian Approach

- Frequentist inference about a statistic (e.g. a regression coef.) is obtained through the assumption that the process generating the data could be repeated a large number of times.

- Specifically, frequentist inference is about the proportion of the time that, in the long-run, realizations of this statistic will fall within some interval.

- If there is no long-run, or possibility of repetition, then probabilistic summaries are not appropriate.

- Suppose we ha the data set $D = \{(y_i, x_i)|i = 1, 2..., n\}$ where x_i is the k variate predictors.

  The response variable y is binary.

$$y_i = \begin{cases} 1 & \text{Success} \\ 0 & \text{Failure} \end{cases}$$

  where $i = 1, 2, ..., n$.

- The binary response model:

$$\begin{aligned} P(y_i = 1|x_i, \beta) &= \Phi(x_i^T \beta) \\ P(y_i = 0|x_i, \beta) &= \Phi(-x_i^T \beta) \end{aligned}$$

- Suppose we ha the data set $D = \{(y_i, x_i)|i = 1, 2..., n\}$ where x_i is the k variate predictors.

  The response variable y is binary.

$$y_i = \begin{cases} 1 & \text{Success} \\ 0 & \text{Failure} \end{cases}$$

  where $i = 1, 2, ..., n$.

- The binary response model:

$$\begin{aligned} P(y_i = 1|x_i, \beta) &= \Phi(x_i^T \beta) \\ P(y_i = 0|x_i, \beta) &= \Phi(-x_i^T \beta) \end{aligned}$$

- Suppose we ha the data set $D = \{(y_i, x_i) | i = 1, 2..., n\}$ where x_i is the k variate predictors.

  The response variable y is binary.

$$y_i = \begin{cases} 1 & \text{Success} \\ 0 & \text{Failure} \end{cases}$$

  where $i = 1, 2, ..., n$.

- The binary response model:

$$\begin{aligned} P(y_i = 1 | x_i, \beta) &= \Phi(x_i^T \beta) \\ P(y_i = 0 | x_i, \beta) &= \Phi(-x_i^T \beta) \end{aligned}$$

Probit Regression: Data Augmentation

- Data Augmention is a very powerful tecnique
- We re-write the Probit regression model in the following way:

$$z_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, ..., n$$

where $\epsilon_i \sim N(0, 1)$

- Now we can say the observed data is being generated as

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

Probit Regression: Data Augmentation

- Data Augmention is a very powerful tecnique

- We re-write the Probit regression model in the following way:

$$z_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, ..., n$$

where $\epsilon_i \sim N(0, 1)$

- Now we can say the observed data is being generated as

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

- Data Augmention is a very powerful tecnique

- We re-write the Probit regression model in the following way:

$$z_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, ..., n$$

where $\epsilon_i \sim N(0,1)$

- Now we can say the observed data is being generated as

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

- Probit Model:

$$z_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, ..., n$$

where $\epsilon_i \sim N(0, 1)$

- Now we can say the observed data is being generated as

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

- Prior:

$$\beta \sim N(\beta_0, \Lambda_0)$$

Probit Regression: Data Augmentation

- Probit Model:

$$z_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, ..., n$$

where $\epsilon_i \sim N(0, 1)$

- Now we can say the observed data is being generated as

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

- **Prior**:

$$\beta \sim N(\beta_0, \Lambda_0)$$

- Conditional posterior distribution:

$$p(\beta|z, y, x) = N(\beta|E(\beta|z, X), \Lambda_n^{-1})$$

where

$$E(\beta|z, X) = (\Lambda_0^{-1} + X^T X)^{-1}(\Lambda_0^{-1}\beta_0 + X^T X\hat{\beta})$$

where $\hat{\beta} = (X^T X)^{-1} X^T z$ and $\Lambda_n = (\Lambda_0^{-1} + X^T X)$

Probit Regression: Data Augmentation

- Data generating process is:
- **Prior**: $\beta \sim N(\beta_0, \Lambda_0)$
- $z_i | x_i, \beta \sim N(x_i^T \beta, 1)$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Data generating process is:
- **Prior**: $\beta \sim N(\beta_0, \Lambda_0)$
- $z_i | x_i, \beta \sim N(x_i^T \beta, 1)$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Probit Regression: Data Augmentation

- Data generating process is:
- **Prior**: $\beta \sim N(\beta_0, \Lambda_0)$
- $z_i | x_i, \beta \sim N(x_i^T \beta, 1)$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Probit Regression: Data Augmentation

- Data generating process is:
- **Prior**: $\beta \sim N(\beta_0, \Lambda_0)$
- $z_i | x_i, \beta \sim N(x_i^T \beta, 1)$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Probit Regression: Gibbs Sampler

1 Choose an initial value $\beta^{(0)}$

2 For $m = 1, 2, ...$

I Sample $z^{(m)}$ as

$$z^{(m)} \sim \begin{cases} TN_{[0,\infty)}(z_i|x_i^T\beta^{(m-1)}, 1) & \text{if } y_i = 1 \\ TN_{(-\infty,0)}(z_i|x_i^T\beta^{(m-1)}, 1) & \text{if } y_i = 0 \end{cases}$$

II

$$\beta^{(m)} \sim N(\beta|E(\beta|z^{(m)}, x), \Lambda_n^{-1})$$

1 Choose an initial value $\beta^{(0)}$

2 For $m = 1, 2, ...$
   I Sample $z^{(m)}$ as

$$z^{(m)} \sim \begin{cases} TN_{[0,\infty)}(z_i | x_i^T \beta^{(m-1)}, 1) & \text{if } y_i = 1 \\ TN_{(-\infty,0)}(z_i | x_i^T \beta^{(m-1)}, 1) & \text{if } y_i = 0 \end{cases}$$

   II

$$\beta^{(m)} \sim N(\beta | E(\beta | z^{(m)}, x), \Lambda_n^{-1})$$

Probit Regression: Gibbs Sampler

1. Choose an initial value $\beta^{(0)}$
2. For $m = 1, 2, ...$
   - I Sample $z^{(m)}$ as

$$z^{(m)} \sim \begin{cases} TN_{[0,\infty)}(z_i | x_i^T \beta^{(m-1)}, 1) & \text{if } y_i = 1 \\ TN_{(-\infty,0)}(z_i | x_i^T \beta^{(m-1)}, 1) & \text{if } y_i = 0 \end{cases}$$

   - II

$$\beta^{(m)} \sim N(\beta | E(\beta | z^{(m)}, x), \Lambda_n^{-1})$$

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$\begin{aligned} z_i = \beta_0 &+ \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

Application

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$\begin{aligned} z_i = \beta_0 \quad &+ \quad \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \quad \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$\begin{aligned} z_i = \beta_0 &+ \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

- Consider ‘birthwt’ dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, 2..., 189$$

$$\begin{aligned} z_i = \beta_0 &+ \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

Posterior Summary

	mean	sd	2.5%	97.5%
Intercept	-0.057	0.263	-0.579	0.456
Black	0.246	0.213	-0.171	0.663
Other	0.299	0.177	-0.051	0.648
Age	-0.032	0.012	-0.055	-0.008
Smoke	0.380	0.170	0.044	0.711

- The 95% CI of the coefficient for race (black and others with respect to white) include 0. Therefore we can say race does not have effect on low birth weight of new born.
- The 95% CI of the coefficient for age doesnot include 0. Negative coefficient indicate the for younder mothers the probabilty of low weight is higher.

Posterior Summary

	mean	sd	2.5%	97.5%
Intercept	-0.057	0.263	-0.579	0.456
Black	0.246	0.213	-0.171	0.663
Other	0.299	0.177	-0.051	0.648
Age	-0.032	0.012	-0.055	-0.008
Smoke	0.380	0.170	0.044	0.711

- The 95% CI of the coefficient for race (black and others with respect to white) include 0. Therefore we can say race does not have effect on low birth weight of new born.

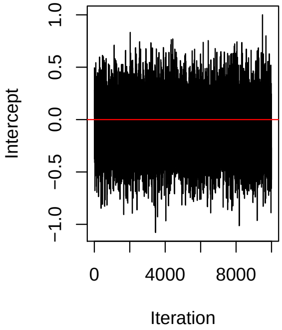
- The 95% CI of the coefficient for age doesnot include 0. Negative coefficient indicate the for younder mothers the probabilty of low weight is higher.

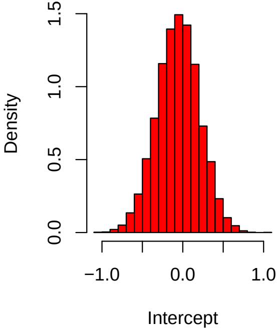
Posterior Summary

	mean	sd	2.5%	97.5%
Intercept	-0.057	0.263	-0.579	0.456
Black	0.246	0.213	-0.171	0.663
Other	0.299	0.177	-0.051	0.648
Age	-0.032	0.012	-0.055	-0.008
Smoke	0.380	0.170	0.044	0.711

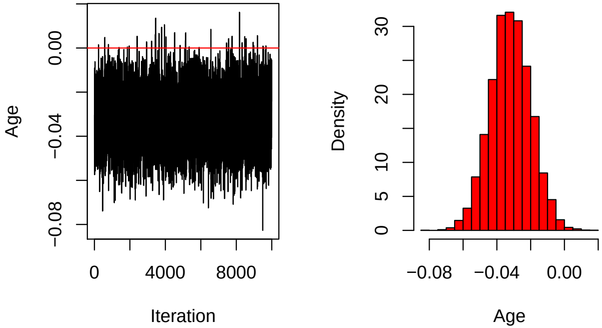
- The 95% CI for the coefficient for smoking status does not include 0. It indicate that if a mother smokes during pregnancy than probability of low birth weight of her child is significantly higher.

Application

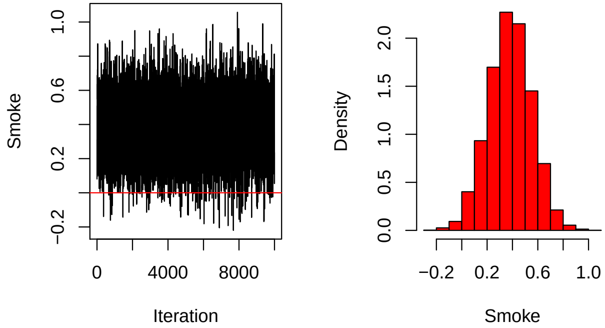




Application



Application



A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: More on MCMC

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Gibbs Sampling
2. Metropolis Hastings Algorithms
3. MCMC Diagnostics

Outline

1. Gibbs Sampling
2. Metropolis Hastings Algorithms
3. MCMC Diagnostics

Outline

1. Gibbs Sampling
2. Metropolis Hastings Algorithms
3. MCMC Diagnostics

What is Gibbs Sampling?

- In the last module we have seen that Gibbs sampling is being introduced in Probit regression model and the implementation becomes really simple ...
- Suppose we have a joint posterior distribution $p(\theta_1, ..., \theta_k|data)$ that we want to sample from.
- We can use the Gibbs sampler to sample from the target distribution if we know the full conditional posterior distributions for each parameter.
- For each parameter, the full conditional posterior distribution is the distribution of the parameter conditional on the known information and all the other parameters: $p(\theta_j|\theta_{-j}, y)$

What is Gibbs Sampling?

- In the last module we have seen that Gibbs sampling is being introduced in Probit regression model and the implementation becomes really simple ...

- Suppose we have a joint posterior distribution $p(\theta_1, ..., \theta_k|data)$ that we want to sample from.

- We can use the Gibbs sampler to sample from the target distribution if we know the full conditional posterior distributions for each parameter.

- For each parameter, the full conditional posterior distribution is the distribution of the parameter conditional on the known information and all the other parameters: $p(\theta_j|\theta_{-j}, y)$

What is Gibbs Sampling?

- In the last module we have seen that Gibbs sampling is being introduced in Probit regression model and the implementation becomes really simple ...

- Suppose we have a joint posterior distribution $p(\theta_1, ..., \theta_k | data)$ that we want to sample from.

- We can use the Gibbs sampler to sample from the target distribution if we know the full conditional posterior distributions for each parameter.

- For each parameter, the full conditional posterior distribution is the distribution of the parameter conditional on the known information and all the other parameters: $p(\theta_j | \theta_{-j}, y)$

What is Gibbs Sampling?

- In the last module we have seen that Gibbs sampling is being introduced in Probit regression model and the implementation becomes really simple ...

- Suppose we have a joint posterior distribution $p(\theta_1, ..., \theta_k|data)$ that we want to sample from.

- We can use the Gibbs sampler to sample from the target distribution if we know the full conditional posterior distributions for each parameter.

- For each parameter, the full conditional posterior distribution is the distribution of the parameter conditional on the known information and all the other parameters: $p(\theta_j|\theta_{-j}, y)$

The Hammersley-Clifford Theorem

- For two blocks - Suppose we have a joint density $f(x, y)$. The theorem proves that we can write out the joint density in terms of the conditional densities. $f(x|y)$ and $f(y|x)$

$$f(x, y) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dy}$$

- We can write the denominator as

$$\begin{aligned} \int \frac{f(y|x)}{f(x|y)} dy &= \int \frac{\frac{f(x,y)}{f(x)}}{\frac{f(x,y)}{f(y)}} dy \\ &= \int \frac{f(y)}{f(x)} dy \\ &= \frac{1}{f(x)} \end{aligned}$$

- For two blocks - Suppose we have a joint density $f(x,y)$. The theorem proves that we can write out the joint density in terms of the conditional densities. $f(x|y)$ and $f(y|x)$

$$f(x,y) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dy}$$

- We can write the denominator as

$$\begin{aligned}
\int \frac{f(y|x)}{f(x|y)} dy &= \int \frac{\frac{f(x,y)}{f(x)}}{\frac{f(x,y)}{f(y)}} dy \\
&= \int \frac{f(y)}{f(x)} dy \\
&= \frac{1}{f(x)}
\end{aligned}$$

- Thus, our right-hand side is

$$
\begin{aligned}
\frac{f(y|x)}{\frac{1}{f(x)}} &= f(y|x)f(x) \\
&= f(x,y)
\end{aligned}
$$

- The theorem shows that knowledge of the conditional densities allows us to get the joint density.

- This works for more than two blocks of parameters.

- But how do we figure out the full conditionals?

- Thus, our right-hand side is

$$\begin{aligned}\frac{f(y|x)}{\frac{1}{f(x)}} &= f(y|x)f(x) \\ &= f(x,y)\end{aligned}$$

- The theorem shows that knowledge of the conditional densities allows us to get the joint density.

- This works for more than two blocks of parameters.

- But how do we figure out the full conditionals?

- Thus, our right-hand side is

$$\frac{f(y|x)}{\frac{1}{f(x)}} = f(y|x)f(x)$$
$$= f(x,y)$$

- The theorem shows that knowledge of the conditional densities allows us to get the joint density.

- This works for more than two blocks of parameters.

- But how do we figure out the full conditionals?

Steps to Calculating Full Conditional Distri

- Target posterior density $p(\theta|y)$
- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

- Target posterior density $p(\theta|y)$
- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

- Target posterior density $p(\theta|y)$

- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

- Target posterior density $p(\theta|y)$
- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

Steps to Calculating Full Conditional Distrib

- Target posterior density $p(\theta|y)$

- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

Steps to Calculating Full Conditional Distri

- Target posterior density $p(\theta|y)$

- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

- Target posterior density $p(\theta|y)$

- To calculate the full conditionals for each θ:
  1. consider full posterior ignoring constant proportion
  2. split the parameter vector into 2, say $\theta = (\theta_1, \theta_2)$
  3. consider a block of parameters (say, θ_1) and ignore everything that doesn't depend on θ_1.
  4. Use the knowledge of distributions to figure out what is the normalizing constant (and thus what is the full conditional distribution $p(\theta_1|\theta_{-1}, y)$)
  5. Repeat the steps for other blocks

You can generalize it for k blocks

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler Steps

- Suppose that we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, θ_3, θ_2, θ_3.

- The steps to a Gibbs Sampler are
  1. Pick a vector of starting values $\theta^{(0)}$ (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)
  2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.
  3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.(Note that we must use the updated value of $\theta_1^{(1)}$.)
  4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

- Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and **P** to get $\Pi^{(1)}$ and then drawing $\theta^{(1)}$ from $\Pi^{(1)}$.

   5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ and continually using the most updated.

   6. Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$. values.
   7. Optional burn-in and/or thinning.

- 'Gibbs sample' is a Markov chain with a bunch of draws of θ that are approximately from our posterior.

Gibbs Sampler Steps

- Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and **P** to get $\Pi^{(1)}$ and then drawing $\theta^{(1)}$ from $\Pi^{(1)}$.
  5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ and continually using the most updated.
  6. Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$. values.
  7. Optional burn-in and/or thinning.
- 'Gibbs sample' is a Markov chain with a bunch of draws of θ that are approximately from our posterior.

Gibbs Sampler Steps

- Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and **P** to get $\Pi^{(1)}$ and then drawing $\theta^{(1)}$ from $\Pi^{(1)}$.

  5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ and continually using the most updated.

  6. Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$. values.
  7. Optional burn-in and/or thinning.

- 'Gibbs sample' is a Markov chain with a bunch of draws of θ that are approximately from our posterior.

Gibbs Sampler Steps

- Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and **P** to get $\Pi^{(1)}$ and then drawing $\theta^{(1)}$ from $\Pi^{(1)}$.
  5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ and continually using the most updated.
  6. Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$. values.
  7. Optional burn-in and/or thinning.
- 'Gibbs sample' is a Markov chain with a bunch of draws of θ that are approximately from our posterior.

- Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and **P** to get $\Pi^{(1)}$ and then drawing $\theta^{(1)}$ from $\Pi^{(1)}$.
  5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ and continually using the most updated.
  6. Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$. values.
  7. Optional burn-in and/or thinning.
- 'Gibbs sample' is a Markov chain with a bunch of draws of θ that are approximately from our posterior.

Application from Robert and Casella,(2004)

- Consider the Poisson-Gamma hierarchical model discussed in Module 7 on hierarchical models

- Suppose we have data of the number of failures (y_i) for each of 10 pumps in a nuclear plant.

- We also have the times (t_i) at which each pump was observed.

```
> y <- c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22)
> t <- c(94, 16, 63, 126, 5, 31, 1, 1, 2, 10)
> rbind(y,t)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
y	5	1	5	14	3	19	1	1	4	22
t	94	16	63	126	5	31	1	1	2	10

- Consider the Poisson-Gamma hierarchical model discussed in Module 7 on hierarchical models

- Suppose we have data of the number of failures (y_i) for each of 10 pumps in a nuclear plant.

- We also have the times (t_i) at which each pump was observed.

```
> y <- c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22)
> t <- c(94, 16, 63, 126, 5, 31, 1, 1, 2, 10)
> rbind(y,t)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
y	5	1	5	14	3	19	1	1	4	22
t	94	16	63	126	5	31	1	1	2	10

- Consider the Poisson-Gamma hierarchical model discussed in Module 7 on hierarchical models

- Suppose we have data of the number of failures (y_i) for each of 10 pumps in a nuclear plant.

- We also have the times (t_i) at which each pump was observed.

```
> y <- c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22)
> t <- c(94, 16, 63, 126, 5, 31, 1, 1, 2, 10)
> rbind(y,t)

  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
y    5    1    5   14    3   19    1    1    4    22
t   94   16   63  126    5   31    1    1    2    10
```

- Consider the Poisson-Gamma hierarchical model discussed in Module 7 on hierarchical models

- Suppose we have data of the number of failures (y_i) for each of 10 pumps in a nuclear plant.

- We also have the times (t_i) at which each pump was observed.

```
> y <- c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22)
> t <- c(94, 16, 63, 126, 5, 31, 1, 1, 2, 10)
> rbind(y,t)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
y    5    1    5   14    3   19    1    1    4    22
t   94   16   63  126    5   31    1    1    2    10
```

Application: Gibbs Sampler for Poisson Gamma Models

- Our objective is to model the number of failures with a Poisson model, where the expected number of failure λ_i differs for each pump.

- Since the observed time for each pump is different; it is required to scale each λ_i by its observed time t_i .

- Poisson likelihood:

$$\prod_{i=1}^{10} Poisson(\lambda_i t_i)$$

- consider $Gamma(\alpha, \beta)$ prior on λ_i with $\alpha = 2$, so λ_i's are drawn from the same distribution.

Application: Gibbs Sampler for Poisson Gamma Models

- Our objective is to model the number of failures with a Poisson model, where the expected number of failure λ_i differs for each pump.

- Since the observed time for each pump is different; it is required to scale each λ_i by its observed time t_i .

- Poisson likelihood:

$$\prod_{i=1}^{10} Poisson(\lambda_i t_i)$$

- consider $Gamma(\alpha, \beta)$ prior on λ_i with $\alpha = 2$, so λ_i's are drawn from the same distribution.

Application: Gibbs Sampler for Poisson Gamma Models

- Our objective is to model the number of failures with a Poisson model, where the expected number of failure λ_i differs for each pump.

- Since the observed time for each pump is different; it is required to scale each λ_i by its observed time t_i .

- Poisson likelihood:

$$\prod_{i=1}^{10} Poisson(\lambda_i t_i)$$

- consider $Gamma(\alpha, \beta)$ prior on λ_i with $\alpha = 2$, so λ_i's are drawn from the same distribution.

- Our objective is to model the number of failures with a Poisson model, where the expected number of failure λ_i differs for each pump.

- Since the observed time for each pump is different; it is required to scale each λ_i by its observed time t_i .

- Poisson likelihood:

$$\prod_{i=1}^{10} Poisson(\lambda_i t_i)$$

- consider $Gamma(\alpha, \beta)$ prior on λ_i with $\alpha = 2$, so λ_i's are drawn from the same distribution.

Application: Gibbs Sampler for Poisson Gamma Models

- Suppose $Gamma(\nu, \delta)$ prior on β with $\nu = 0.01$ and $\delta = 1$
- So our model has 11 parameters that are unknown (10 λ_i s and β).
- Full posterior is

$$p(\lambda, \beta | \mathbf{y}, \mathbf{t}) \propto \left(\prod_{i=1}^{10} Poisson(\lambda_i t_i) \times Gamma(\alpha, \beta) \right) \times Gamma(\nu, \delta)$$

Application: Gibbs Sampler for Poisson Gamma Models

- Suppose $Gamma(\nu, \delta)$ prior on β with $\nu = 0.01$ and $\delta = 1$
- So our model has 11 parameters that are unknown (10 λ_i s and β).
- Full posterior is

$$p(\lambda, \beta|\mathbf{y}, \mathbf{t}) \propto \left(\prod_{i=1}^{10} Poisson(\lambda_i t_i) \times Gamma(\alpha, \beta)\right) \times Gamma(\nu, \delta)$$

Application: Gibbs Sampler for Poisson Gamma Models

- Suppose $Gamma(\nu, \delta)$ prior on β with $\nu = 0.01$ and $\delta = 1$
- So our model has 11 parameters that are unknown (10 λ_i s and β).
- Full posterior is

$$p(\lambda, \beta|\mathbf{y}, \mathbf{t}) \propto \left(\prod_{i=1}^{10} Poisson(\lambda_i t_i) \times Gamma(\alpha, \beta) \right) \times Gamma(\nu, \delta)$$

Application: Gibbs Sampler for Poisson Gamma Models

$$
\begin{aligned}
p(\lambda, \beta | \mathbf{y}, \mathbf{t}) &\propto \left(\prod_{i=1}^{10} Poisson(\lambda_i t_i) \times Gamma(\alpha, \beta) \right) \\
&\quad \times Gamma(\nu, \delta) \\
&= \left(\prod_{i=1}^{10} e^{-\lambda_i t_i} \frac{(\lambda_i t_i)^{y_i}}{y_i!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha_i - 1} e^{-\beta \lambda_i} \right) \\
&\quad \times \frac{\delta^\nu}{\Gamma(\nu)} \beta^{\nu-1} e^{-\delta\beta} \\
&\propto \left(\prod_{i=1}^{10} e^{-\lambda_i t_i} (\lambda_i t_i)^{y_i} \times \beta^\alpha \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \right) \times \beta^{\nu-1} e^{-\delta\beta} \\
&\propto \left(\prod_{i=1}^{10} \lambda_i^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i} \right) \beta^{10\alpha+\nu-1} e^{-\delta\beta}
\end{aligned}
$$

Application: Gibbs Sampler for Poisson Gamma Models

- Full conditional posterior distributions:
  1. $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t}) \propto \lambda_i^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i}$

     $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t})$ is $Gamma(y_i + \alpha, t_i + \beta)$
  2. $p(\beta|\lambda, \mathbf{y}, \mathbf{t}) \propto e^{-\beta(\delta+\sum_{i=1}^{10}\lambda_i)} \beta^{10\alpha+\nu-1}$

     $p(\beta|\lambda, \mathbf{y}, \mathbf{t})$ is $Gamma(10\alpha + \nu, \delta + \sum_{i=1}^{10}\lambda_i)$

Application: Gibbs Sampler for Poisson Gamma Models

- Full conditional posterior distributions:
  1. $p(\lambda_i|\lambda_{-i},\beta,\mathbf{y},\mathbf{t}) \propto \lambda_i^{y_i+\alpha-1}e^{-(t_i+\beta)\lambda_i}$

     $p(\lambda_i|\lambda_{-i},\beta,\mathbf{y},\mathbf{t})$ is $Gamma(y_i+\alpha, t_i+\beta)$

  2. $p(\beta|\lambda,\mathbf{y},\mathbf{t}) \propto e^{-\beta(\delta+\sum_{i=1}^{10}\lambda_i)}\beta^{10\alpha+\nu-1}$

     $p(\beta|\lambda,\mathbf{y},\mathbf{t})$ is $Gamma(10\alpha+\nu, \delta+\sum_{i=1}^{10}\lambda_i)$

Application: Gibbs Sampler for Poisson Gamma Models

- Full conditional posterior distributions:
  1. $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t}) \propto \lambda_i^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i}$

     $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t})$ is $Gamma(y_i+\alpha, t_i+\beta)$
  2. $p(\beta|\lambda, \mathbf{y}, \mathbf{t}) \propto e^{-\beta(\delta+\sum_{i=1}^{10}\lambda_i)}\beta^{10\alpha+\nu-1}$

     $p(\beta|\lambda, \mathbf{y}, \mathbf{t})$ is $Gamma(10\alpha+\nu, \delta+\sum_{i=1}^{10}\lambda_i)$

Application: Gibbs Sampler for Poisson Gamma Models

- Full conditional posterior distributions:
  1. $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t}) \propto \lambda_i^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i}$

     $p(\lambda_i|\lambda_{-i}, \beta, \mathbf{y}, \mathbf{t})$ is $Gamma(y_i + \alpha, t_i + \beta)$
  2. $p(\beta|\lambda, \mathbf{y}, \mathbf{t}) \propto e^{-\beta(\delta+\sum_{i=1}^{10}\lambda_i)}\beta^{10\alpha+\nu-1}$

     $p(\beta|\lambda, \mathbf{y}, \mathbf{t})$ is $Gamma(10\alpha + \nu, \delta + \sum_{i=1}^{10}\lambda_i)$

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                                , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                           , delta + sum(lambda))
+ }
```

Coding the Gibbs Sampler

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                           , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                      , delta + sum(lambda))
+ }
```

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                              , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                         , delta + sum(lambda))
+ }
```

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                                  , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                         , delta + sum(lambda))
+ }
```

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                           , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                     , delta + sum(lambda))
+ }
```

1. Define starting values of $\beta^{(0)}$

```
> beta.cur<-1
```

2. Draw $\lambda^{(1)}$ from its full conditional (we draw all the λ_i's as a block because they all only depend on β and not each other)

```
> lambda.update <- function(alpha, beta, y, t) {
+       rgamma(length(y), y + alpha, t + beta)
+ }
```

3. Draw $\beta^{(1)}$ from its full conditional, using $\lambda^{(1)}$

```
> beta.update <- function(alpha, gamma, delta
+                               , lambda, y) {
+       rgamma(1, length(y) * alpha + gamma
+                         , delta + sum(lambda))
+ }
```

4. Repeat using most updated values until we get M draws.
5. Optional burn-in and thinning.
6. Write it as a function so that you can use it repeatedly.

4. Repeat using most updated values until we get M draws.

5. Optional burn-in and thinning.

6. Write it as a function so that you can use it repeatedly.

4. Repeat using most updated values until we get M draws.
5. Optional burn-in and thinning.
6. Write it as a function so that you can use it repeatedly.

7. Do Monte Carlo Integration on the resulting Markov chain, which are samples approximately from the posterior.

```
> round(colMeans(posterior$lambda.draws),digit=2)
 [1] 0.07 0.16 0.11 0.12 0.65 0.62 0.85 0.86 1.31 1.90
> round(mean(posterior$beta.draws),digit=2)
[1] 2.68
> round(apply(posterior$lambda.draws, 2, sd),digit=2)
 [1] 0.03 0.09 0.04 0.03 0.30 0.14 0.54 0.54 0.58 0.40
> round(sd(posterior$beta.draws),digit=2)
[1] 0.75
```

- Suppose we have a posterior $p(\theta|y)$ that we want to sample from, but
  - the posterior does not look like any known distribution that we know
  - If the posterior consists of more than 2 parameters then grid approximations is not possible.
  - some (or all) of the full conditionals do not follow known distributions. Therefore Gibbs sampling is not possible for those whose full conditionals are not known.
- In such cases, we can use the **Metropolis-Hastings** algorithm.

- Suppose we have a posterior $p(\theta|y)$ that we want to sample from, but
  - the posterior does not look like any known distribution that we know
  - If the posterior consists of more than 2 parameters then grid approximations is not possible.
  - some (or all) of the full conditionals do not follow known distributions. Therefore Gibbs sampling is not possible for those whose full conditionals are not known.
- In such cases, we can use the **Metropolis-Hastings** algorithm.

The Metropolis-Hastings Algorithm

- Suppose we have a posterior $p(\theta|y)$ that we want to sample from, but
  - the posterior does not look like any known distribution that we know
  - If the posterior consists of more than 2 parameters then grid approximations is not possible.
  - some (or all) of the full conditionals do not follow known distributions. Therefore Gibbs sampling is not possible for those whose full conditionals are not known.
- In such cases, we can use the **Metropolis-Hastings** algorithm.

- Suppose we have a posterior $p(\theta|y)$ that we want to sample from, but
  - the posterior does not look like any known distribution that we know
  - If the posterior consists of more than 2 parameters then grid approximations is not possible.
  - some (or all) of the full conditionals do not follow known distributions. Therefore Gibbs sampling is not possible for those whose full conditionals are not known.
- In such cases, we can use the **Metropolis-Hastings** algorithm.

- Suppose we have a posterior $p(\theta|y)$ that we want to sample from, but
  - the posterior does not look like any known distribution that we know
  - If the posterior consists of more than 2 parameters then grid approximations is not possible.
  - some (or all) of the full conditionals do not follow known distributions. Therefore Gibbs sampling is not possible for those whose full conditionals are not known.
- In such cases, we can use the **Metropolis-Hastings** algorithm.

The Metropolis-Hastings Algorithm :

1. Choose starting value $^{(0)}$
2. At iteration t, draw a candidate θ^* from a proposal distribution $q_t(\theta^*|\theta^{t-1}))$.
3. Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

The Metropolis-Hastings Algorithm :

1. Choose starting value ${}^{(0)}$

2. At iteration t, draw a candidate θ^* from a proposal distribution $q_t(\theta^*|\theta^{t-1}))$.

3. Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

The Metropolis-Hastings Algorithm :

1. Choose starting value $^{(0)}$
2. At iteration t, draw a candidate θ^* from a proposal distribution $q_t(\theta^*|\theta^{t-1}))$.
3. Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

4. Accept $\theta^{(t)} = \theta^*$ if $U < \min(1, r)$ where $U \sim uniform(0, 1)$
   Otherwise $\theta^{(t)} = \theta^{(t-1)}$

5. Repeat steps 2-4 M times to get M draws from $p(\theta|y)$ with optional burn-in and/or thinning

4. Accept $\theta^{(t)} = \theta^*$ if $U < \min(1, r)$ where $U \sim uniform(0, 1)$
   Otherwise $\theta^{(t)} = \theta^{(t-1)}$

5. Repeat steps 2-4 M times to get M draws from $p(\theta|y)$ with optional burn-in and/or thinning

Step 1: Choose a starting value $\theta^{(0)}$

- This is similar to draw sample from initial stationary distribution.
- Note that $\theta^{(0)}$ must have positive probability.

$$p(\theta^{(0)}|y) > 0$$

Step 1: Choose a starting value $\theta^{(0)}$

- This is similar to draw sample from initial stationary distribution.
- Note that $\theta^{(0)}$ must have positive probability.

$$p(\theta^{(0)}|y) > 0$$

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- The proposal distribution $q_t(\theta^*|\theta^{(t-1)})$ determines where it moves in the next iteration of the Markov chain.
- The proposal distribution is analogous to transition kernel.
- The support of proposal density must contain the support of the posterior
- If a symmetric proposal distribution that is dependent on $\theta^{(t-1)}$, then it is known as random walk Metropolis sampling.

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- The proposal distribution $q_t(\theta^*|\theta^{(t-1)})$ determines where it moves in the next iteration of the Markov chain.
- The proposal distribution is analogous to transition kernel.
- The support of proposal density must contain the support of the posterior
- If a symmetric proposal distribution that is dependent on $\theta^{(t-1)}$, then it is known as random walk Metropolis sampling.

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- The proposal distribution $q_t(\theta^*|\theta^{(t-1)})$ determines where it moves in the next iteration of the Markov chain.
- The proposal distribution is analogous to transition kernel.
- The support of proposal density must contain the support of the posterior
- If a symmetric proposal distribution that is dependent on $\theta^{(t-1)}$, then it is known as random walk Metropolis sampling.

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- The proposal distribution $q_t(\theta^*|\theta^{(t-1)})$ determines where it moves in the next iteration of the Markov chain.
- The proposal distribution is analogous to transition kernel.
- The support of proposal density must contain the support of the posterior
- If a symmetric proposal distribution that is dependent on $\theta^{(t-1)}$, then it is known as random walk Metropolis sampling.

- If the proposal distribution does not depend on $\theta^{(t-1)}$

$$q_t(\theta^*|\theta^{(t-1)}) = q_t(\theta^*)$$

then it is known as independent Metropolis-Hasting sampling

- In this case all candidate draws θ^* are drawn from the same distribution, regardless of where the previous draw was.

- This can be very efficient or very inefficient, depending how close the proposal density with respect to posterior density.

- Typically, if proposal density has heavier tails than the posterior, then chain will behave nicely

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- If the proposal distribution does not depend on $\theta^{(t-1)}$

$$q_t(\theta^*|\theta^{(t-1)}) = q_t(\theta^*)$$

then it is known as independent Metropolis-Hasting sampling

- In this case all candidate draws θ^* are drawn from the same distribution, regardless of where the previous draw was.

- This can be very efficient or very inefficient, depending how close the proposal density with respect to posterior density.

- Typically, if proposal density has heavier tails than the posterior, then chain will behave nicely

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- If the proposal distribution does not depend on $\theta^{(t-1)}$

$$q_t(\theta^*|\theta^{(t-1)}) = q_t(\theta^*)$$

  then it is known as independent Metropolis-Hasting sampling

- In this case all candidate draws θ^* are drawn from the same distribution, regardless of where the previous draw was.

- This can be very efficient or very inefficient, depending how close the proposal density with respect to posterior density.

- Typically, if proposal density has heavier tails than the posterior, then chain will behave nicely

Step 2: Draw $\theta^{(*)}$ from $q_t(\theta^*|\theta^{(t-1)})$

- If the proposal distribution does not depend on $\theta^{(t-1)}$

$$q_t(\theta^*|\theta^{(t-1)}) = q_t(\theta^*)$$

then it is known as independent Metropolis-Hasting sampling

- In this case all candidate draws θ^* are drawn from the same distribution, regardless of where the previous draw was.
- This can be very efficient or very inefficient, depending how close the proposal density with respect to posterior density.
- Typically, if proposal density has heavier tails than the posterior, then chain will behave nicely

Step 3: Compute acceptance ratio r .

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

- In the case where our proposal density is symmetric

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$$

- If the candidate sample has higher probability than the current sample, then the candidate is better so we definitely accept it.
- Otherwise, the candidate is accepted according to the ratio of the probabilities of the candidate and current samples.
- Since r is a ratio, we only need to know the $p(\theta|y)$ upto a constant.

Step 3: Compute acceptance ratio r .

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

- In the case where our proposal density is symmetric

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$$

- If the candidate sample has higher probability than the current sample, then the candidate is better so we definitely accept it.

- Otherwise, the candidate is accepted according to the ratio of the probabilities of the candidate and current samples.

- Since r is a ratio, we only need to know the $p(\theta|y)$ upto a constant.

Step 3: Compute acceptance ratio r .

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

- In the case where our proposal density is symmetric

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$$

- If the candidate sample has higher probability than the current sample, then the candidate is better so we definitely accept it.
- Otherwise, the candidate is accepted according to the ratio of the probabilities of the candidate and current samples.
- Since r is a ratio, we only need to know the $p(\theta|y)$ upto a constant.

Step 3: Compute acceptance ratio r .

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

- In the case where our proposal density is symmetric

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$$

- If the candidate sample has higher probability than the current sample, then the candidate is better so we definitely accept it.
- Otherwise, the candidate is accepted according to the ratio of the probabilities of the candidate and current samples.
- Since r is a ratio, we only need to know the $p(\theta|y)$ upto a constant.

Step 3: Compute acceptance ratio r .

$$r = \frac{p(\theta^*|y)/q_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/q_t(\theta^{(t-1)}|\theta^*)}$$

- In the case where our proposal density is symmetric

$$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$$

- If the candidate sample has higher probability than the current sample, then the candidate is better so we definitely accept it.
- Otherwise, the candidate is accepted according to the ratio of the probabilities of the candidate and current samples.
- Since r is a ratio, we only need to know the $p(\theta|y)$ upto a constant.

Step 4: Decide to whether accept θ^*

- Accept θ^* as $\theta^{(t)}$ with probability $min(r, 1)$
- For each θ^*, draw a value u from the $Uniform(0, 1)$ distribution.
- If $u \leq r$ accept θ^* as $\theta^{(t)}$. Otherwise $\theta^{(t-1)} = \theta^{(t)}$
- Candidate sample with higher density than the current samples are always accepted
- Unlike in rejection sampling, each iteration always produces a sample, either θ^* or $\theta^{(t-1)}$.

Step 4: Decide to whether accept θ^*

- Accept θ^* as $\theta^{(t)}$ with probability $min(r, 1)$
- For each θ^*, draw a value u from the $Uniform(0, 1)$ distribution.
- If $u \leq r$ accept θ^* as $\theta^{(t)}$. Otherwise $\theta^{(t-1)} = \theta^{(t)}$
- Candidate sample with higher density than the current samples are always accepted
- Unlike in rejection sampling, each iteration always produces a sample, either θ^* or $\theta^{(t-1)}$.

Step 4: Decide to whether accept θ^*

- Accept θ^* as $\theta^{(t)}$ with probability $min(r, 1)$
- For each θ^*, draw a value u from the $Uniform(0, 1)$ distribution.
- If $u \leq r$ accept θ^* as $\theta^{(t)}$. Otherwise $\theta^{(t-1)} = \theta^{(t)}$
- Candidate sample with higher density than the current samples are always accepted
- Unlike in rejection sampling, each iteration always produces a sample, either θ^* or $\theta^{(t-1)}$.

Step 4: Decide to whether accept θ^*

- Accept θ^* as $\theta^{(t)}$ with probability $min(r, 1)$
- For each θ^*, draw a value u from the $Uniform(0, 1)$ distribution.
- If $u \leq r$ accept θ^* as $\theta^{(t)}$. Otherwise $\theta^{(t-1)} = \theta^{(t)}$
- Candidate sample with higher density than the current samples are always accepted
- Unlike in rejection sampling, each iteration always produces a sample, either θ^* or $\theta^{(t-1)}$.

Step 4: Decide to whether accept θ^*

- Accept θ^* as $\theta^{(t)}$ with probability $min(r, 1)$
- For each θ^*, draw a value u from the $Uniform(0, 1)$ distribution.
- If $u \leq r$ accept θ^* as $\theta^{(t)}$. Otherwise $\theta^{(t-1)} = \theta^{(t)}$
- Candidate sample with higher density than the current samples are always accepted
- Unlike in rejection sampling, each iteration always produces a sample, either θ^* or $\theta^{(t-1)}$.

Acceptance Rate

- It is important to monitor the *acceptance rate* (the fraction of candidate samples that are accepted) of your Metropolis-Hastings algorithm.

- If acceptance rate is too high, the chain is probably not mixing well. That is the chain is not moving around the parameter space quickly enough.

- If the acceptance rate is too low, your algorithm is too inefficient, that is rejecting too many candidate samples.

- What is too high and too low depends on the specific algorithm, but generally
  - **random walk**: somewhere between 0.25 and 0.50 is recommended
  - **independent**: something close to 1 is preferred

- It is important to monitor the *acceptance rate* (the fraction of candidate samples that are accepted) of your Metropolis-Hastings algorithm.
- If acceptance rate is too high, the chain is probably not mixing well. That is the chain is not moving around the parameter space quickly enough.
- If the acceptance rate is too low, your algorithm is too inefficient, that is rejecting too many candidate samples.
- What is too high and too low depends on the specific algorithm, but generally
  - **random walk**: somewhere between 0.25 and 0.50 is recommended
  - **independent**: something close to 1 is preferred

Acceptance Rate

- It is important to monitor the *acceptance rate* (the fraction of candidate samples that are accepted) of your Metropolis-Hastings algorithm.
- If acceptance rate is too high, the chain is probably not mixing well. That is the chain is not moving around the parameter space quickly enough.
- If the acceptance rate is too low, your algorithm is too inefficient, that is rejecting too many candidate samples.
- What is too high and too low depends on the specific algorithm, but generally
  - **random walk**: somewhere between 0.25 and 0.50 is recommended
  - **independent**: something close to 1 is preferred

Acceptance Rate

- It is important to monitor the *acceptance rate* (the fraction of candidate samples that are accepted) of your Metropolis-Hastings algorithm.
- If acceptance rate is too high, the chain is probably not mixing well. That is the chain is not moving around the parameter space quickly enough.
- If the acceptance rate is too low, your algorithm is too inefficient, that is rejecting too many candidate samples.
- What is too high and too low depends on the specific algorithm, but generally
  - **random walk**: somewhere between 0.25 and 0.50 is recommended
  - **independent**: something close to 1 is preferred

Acceptance Rate

- It is important to monitor the *acceptance rate* (the fraction of candidate samples that are accepted) of your Metropolis-Hastings algorithm.
- If acceptance rate is too high, the chain is probably not mixing well. That is the chain is not moving around the parameter space quickly enough.
- If the acceptance rate is too low, your algorithm is too inefficient, that is rejecting too many candidate samples.
- What is too high and too low depends on the specific algorithm, but generally
  - **random walk**: somewhere between 0.25 and 0.50 is recommended
  - **independent**: something close to 1 is preferred

MCMC Convergence

- After the model has converged, samples from the conditional distributions are used to summarize the posterior distribution of parameters of interest θ.
- **Convergence** refers to the idea that eventually the Gibbs Sampler or other MCMC technique that we choose will eventually reach a stationary distribution.
- From this point onwards it stays in this distribution and moved about or "mixes" throughout the subspace forever.

MCMC Convergence

- After the model has converged, samples from the conditional distributions are used to summarize the posterior distribution of parameters of interest θ.

- **Convergence** refers to the idea that eventually the Gibbs Sampler or other MCMC technique that we choose will eventually reach a stationary distribution.

- From this point onwards it stays in this distribution and moved about or "mixes" throughout the subspace forever.

- After the model has converged, samples from the conditional distributions are used to summarize the posterior distribution of parameters of interest θ.
- **Convergence** refers to the idea that eventually the Gibbs Sampler or other MCMC technique that we choose will eventually reach a stationary distribution.
- From this point onwards it stays in this distribution and moved about or "mixes" throughout the subspace forever.

MCMC Convergence

- The general questions for us:
  1. At what point do we know that have we converged to the stationary distribution? (i.e. how long should our "burn-in" period be?
  2. After we have reached the stationary distribution, how many iterations will it take to summarize the posterior distribution?
- The answers to both of these questions remain a bit *ad hoc* because the desirable results that we depend on are only true asymptotically, and we donot want to wait for an infinite number of draws.

MCMC Convergence

- The general questions for us:
  1. At what point do we know that have we converged to the stationary distribution? (i.e. how long should our "burn-in" period be?
  2. After we have reached the stationary distribution, how many iterations will it take to summarize the posterior distribution?
- The answers to both of these questions remain a bit *ad hoc* because the desirable results that we depend on are only true asymptotically, and we donot want to wait for an infinite number of draws.

MCMC Convergence

- The general questions for us:
  1. At what point do we know that have we converged to the stationary distribution? (i.e. how long should our "burn-in" period be?
  2. After we have reached the stationary distribution, how many iterations will it take to summarize the posterior distribution?
- The answers to both of these questions remain a bit *ad hoc* because the desirable results that we depend on are only true asymptotically, and we donot want to wait for an infinite number of draws.

- The general questions for us:
  1. At what point do we know that have we converged to the stationary distribution? (i.e. how long should our "burn-in" period be?
  2. After we have reached the stationary distribution, how many iterations will it take to summarize the posterior distribution?
- The answers to both of these questions remain a bit *ad hoc* because the desirable results that we depend on are only true asymptotically, and we donot want to wait for an infinite number of draws.

Possible problems with MCMC

- The assumed model may not be realistic from a substantive point of view or may not fit.

- Errors in calculation or programming!
  - Often, simple syntax mistakes may be responsible; however, it is possible that the algorithm may not converge to a proper distribution.

- **Slow convergence**: this is the problem we are most likely to run into. The simulation can remain for many iterations in a region heavily influenced by the starting distribution. If the iterations are used to summarize the target distribution, they can yield falsely precise estimates.

this will be the focus of our discussion.

Possible problems with MCMC

- The assumed model may not be realistic from a substantive point of view or may not fit.

- Errors in calculation or programming!
  - Often, simple syntax mistakes may be responsible; however, it is possible that the algorithm may not converge to a proper distribution.

- **Slow convergence**: this is the problem we are most likely to run into. The simulation can remain for many iterations in a region heavily influenced by the starting distribution. If the iterations are used to summarize the target distribution, they can yield falsely precise estimates.

this will be the focus of our discussion.

Possible problems with MCMC

- The assumed model may not be realistic from a substantive point of view or may not fit.

- Errors in calculation or programming!
  - Often, simple syntax mistakes may be responsible; however, it is possible that the algorithm may not converge to a proper distribution.

- **Slow convergence**: this is the problem we are most likely to run into. The simulation can remain for many iterations in a region heavily influenced by the starting distribution. If the iterations are used to summarize the target distribution, they can yield falsely precise estimates.

this will be the focus of our discussion.

Possible problems with MCMC

- The assumed model may not be realistic from a substantive point of view or may not fit.

- Errors in calculation or programming!
  - Often, simple syntax mistakes may be responsible; however, it is possible that the algorithm may not converge to a proper distribution.

- **Slow convergence**: this is the problem we are most likely to run into. The simulation can remain for many iterations in a region heavily influenced by the starting distribution. If the iterations are used to summarize the target distribution, they can yield falsely precise estimates.

this will be the focus of our discussion.

Traceplots

- One intuitive and easily implemented diagnostic tool is a traceplot (or history plot) which plots the parameter value at time t against the iteration number.

- If the model has converged, the traceplot will move snake around the mode of the distribution.

- A clear sign of non-convergence with a traceplot occurs when we observe some trending in the sample space.

- The problem with traceplots is that it may appear that we have converged, however, the chain trapped (for a finite time) in a local region rather exploring the full posterior.

Traceplots

- One intuitive and easily implemented diagnostic tool is a traceplot (or history plot) which plots the parameter value at time t against the iteration number.

- If the model has converged, the traceplot will move snake around the mode of the distribution.

- A clear sign of non-convergence with a traceplot occurs when we observe some trending in the sample space.

- The problem with traceplots is that it may appear that we have converged, however, the chain trapped (for a finite time) in a local region rather exploring the full posterior.

Traceplots

- One intuitive and easily implemented diagnostic tool is a traceplot (or history plot) which plots the parameter value at time t against the iteration number.

- If the model has converged, the traceplot will move snake around the mode of the distribution.

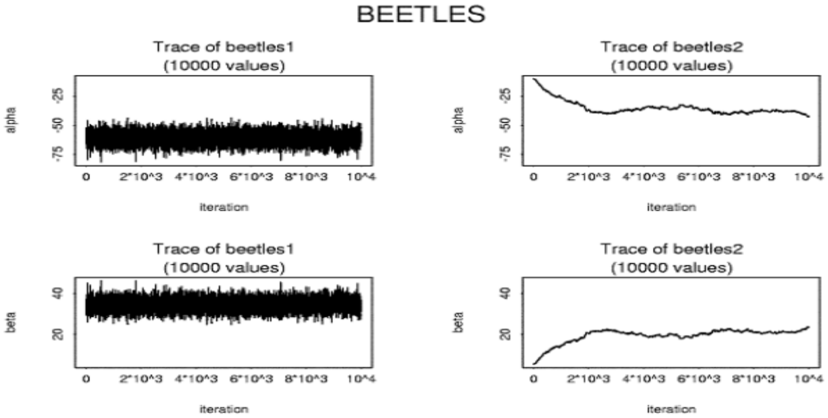
- A clear sign of non-convergence with a traceplot occurs when we observe some trending in the sample space.

- The problem with traceplots is that it may appear that we have converged, however, the chain trapped (for a finite time) in a local region rather exploring the full posterior.

Traceplots

- One intuitive and easily implemented diagnostic tool is a traceplot (or history plot) which plots the parameter value at time t against the iteration number.
- If the model has converged, the traceplot will move snake around the mode of the distribution.
- A clear sign of non-convergence with a traceplot occurs when we observe some trending in the sample space.
- The problem with traceplots is that it may appear that we have converged, however, the chain trapped (for a finite time) in a local region rather exploring the full posterior.

Examples of Apparent Convergence and Non-Convergence Based on a trace plot



- **Autocorrelation** refers to a pattern of serial correlation in the chain, where sequential draws of a parameter, say θ_1, from the conditional distribution are correlated.
- The cause of autocorrelation is that the parameters in our model may be highly correlated, so the Gibbs Sampler will be slow to explore the entire posterior distribution.
- The reason why autocorrelation is important is that it will take a very long time to explore the entire posterior distribution.

Autocorrelation Diagnostic

- **Autocorrelation** refers to a pattern of serial correlation in the chain, where sequential draws of a parameter, say θ_1, from the conditional distribution are correlated.
- The cause of autocorrelation is that the parameters in our model may be highly correlated, so the Gibbs Sampler will be slow to explore the entire posterior distribution.
- The reason why autocorrelation is important is that it will take a very long time to explore the entire posterior distribution.

Autocorrelation Diagnostic

- **Autocorrelation** refers to a pattern of serial correlation in the chain, where sequential draws of a parameter, say θ_1, from the conditional distribution are correlated.
- The cause of autocorrelation is that the parameters in our model may be highly correlated, so the Gibbs Sampler will be slow to explore the entire posterior distribution.
- The reason why autocorrelation is important is that it will take a very long time to explore the entire posterior distribution.

Autocorrelation Diagnostic

- **Autocorrelation** refers to a pattern of serial correlation in the chain, where sequential draws of a parameter, say θ_1, from the conditional distribution are correlated.

- The cause of autocorrelation is that the parameters in our model may be highly correlated, so the Gibbs Sampler will be slow to explore the entire posterior distribution.

- The reason why autocorrelation is important is that it will take a very long time to explore the entire posterior distribution.

Autocorrelation Diagnostic

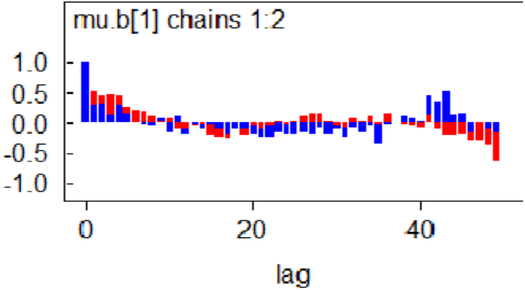
- Note that if the level of autocorrelation is high for a parameter of interest, then a traceplot will be a poor diagnostic for convergence.
- Typically, the level of autocorrelation will decline with an increasing number of lags in the chain (e.g. as we go from the 1000th to the 1010th lags, the level of autocorrelation will often decline.) When this dampening doesn't occur, then you have a problem and will probably want to re-parameterize the model (more on this below).

Autocorrelation Diagnostic

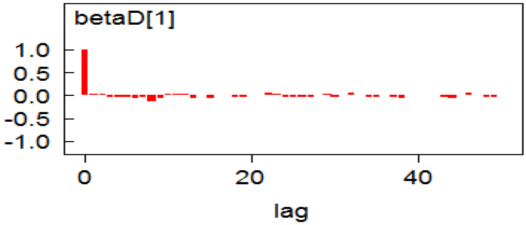
- Note that if the level of autocorrelation is high for a parameter of interest, then a traceplot will be a poor diagnostic for convergence.

- Typically, the level of autocorrelation will decline with an increasing number of lags in the chain (e.g. as we go from the 1000th to the 1010th lags, the level of autocorrelation will often decline.) When this dampening doesn't occur, then you have a problem and will probably want to re-parameterize the model (more on this below).

Example of Model with autocorrelation



Example of Model without autocorrelation



Running Means of parameters

- **Running means**: Once you have taken enough draws to summarize the posterior distribution, then if the model has converged, further samples from a parameter's posterior distribution should not influence the calculation of the mean.

- A plot of the average draw from the conditional distribution of draws 1 through t against t is useful for identifying convergence.

- Note: that you could probably get the same effect with a traceplot.

- **Running means**: Once you have taken enough draws to summarize the posterior distribution, then if the model has converged, further samples from a parameter's posterior distribution should not influence the calculation of the mean.

- A plot of the average draw from the conditional distribution of draws 1 through t against t is useful for identifying convergence.

- Note: that you could probably get the same effect with a traceplot.

Running Means of parameters

- **Running means**: Once you have taken enough draws to summarize the posterior distribution, then if the model has converged, further samples from a parameter's posterior distribution should not influence the calculation of the mean.
- A plot of the average draw from the conditional distribution of draws 1 through t against t is useful for identifying convergence.
- Note: that you could probably get the same effect with a traceplot.

- Kernel density plots (a.k.a. smoothed density; histograms) may be useful diagnostic.

- Sometimes non-convergence is reflected in multimodal distributions. This is especially true if the kernel density plot isn't just multi-modal, but lumpy (you will know what I mean when you see it).

- When you get a lumpy posterior, it may be important to let the algorithm run a bit longer. Often, doing this will allow you to get a much more reasonable summary of the posterior.

- Kernel density plots (a.k.a. smoothed density; histograms) may be useful diagnostic.

- Sometimes non-convergence is reflected in multimodal distributions. This is especially true if the kernel density plot isn't just multi-modal, but lumpy (you will know what I mean when you see it).

- When you get a lumpy posterior, it may be important to let the algorithm run a bit longer. Often, doing this will allow you to get a much more reasonable summary of the posterior.

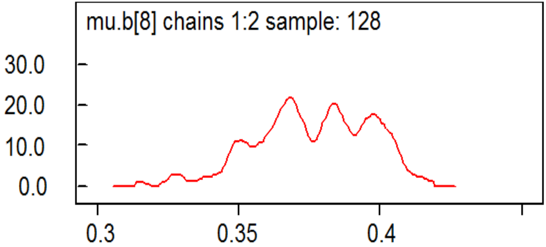
Kernel Density Plot

- Kernel density plots (a.k.a. smoothed density; histograms) may be useful diagnostic.

- Sometimes non-convergence is reflected in multimodal distributions. This is especially true if the kernel density plot isn't just multi-modal, but lumpy (you will know what I mean when you see it).

- When you get a lumpy posterior, it may be important to let the algorithm run a bit longer. Often, doing this will allow you to get a much more reasonable summary of the posterior.

Example of a problematic kernel density plot



A more satisfactory kernel density plot would look more bell-shaped, though it need not be symmetric

- **The Geweke Time-Series Diagnostic**: if a model has converged, then if simulate a large number of draws, the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.

- Technically, this statistic is based on spectral density functions that are beyond the purview of this course and WinBugs does not estimate this statistic directly, but if you export the CODA chain to R the programs CODA and BOA report the Geweke statistic.

- However, a perfectly reasonable way to proceed is look to see whether the posterior means (and variances) of your parameters are approximately the same for different halves of your simulated chain.

- **The Geweke Time-Series Diagnostic**: if a model has converged, then if simulate a large number of draws, the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.

- Technically, this statistic is based on spectral density functions that are beyond the purview of this course and WinBugs does not estimate this statistic directly, but if you export the CODA chain to R the programs CODA and BOA report the Geweke statistic.

- However, a perfectly reasonable way to proceed is look to see whether the posterior means (and variances) of your parameters are approximately the same for different halves of your simulated chain.

Geweke Time-Series Diagnostic

- **The Geweke Time-Series Diagnostic**: if a model has converged, then if simulate a large number of draws, the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.

- Technically, this statistic is based on spectral density functions that are beyond the purview of this course and WinBugs does not estimate this statistic directly, but if you export the CODA chain to R the programs CODA and BOA report the Geweke statistic.

- However, a perfectly reasonable way to proceed is look to see whether the posterior means (and variances) of your parameters are approximately the same for different halves of your simulated chain.

Geweke Time-Series Diagnostic

- **The Geweke Time-Series Diagnostic**: if a model has converged, then if simulate a large number of draws, the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.

- The value of this approach is that by allowing the algorithm to run for a very long time, it may reach areas of the posterior distribution that may not otherwise be reached.

- **The Geweke Time-Series Diagnostic**: if a model has converged, then if simulate a large number of draws, the mean (and variance) of a parameter's posterior distribution from the first half of the chain will be equal to the mean (and variance) from the second half of the chain.

- The value of this approach is that by allowing the algorithm to run for a very long time, it may reach areas of the posterior distribution that may not otherwise be reached.

Gelman-Rubin Diagnostic

- Gelman (especially) argues that the best way to identify non-convergence is to simulate multiple sequences for over-dispersed starting points.

- The intuition is that the behavior of all of the chains should be basically the same.

- Or, as Gelman and Rubin put it, the variance within the chains should be the same as the variance across the chains.

- I think this can be diagnosed pretty easily through traceplots of multiple chains. You want to see if it looks like the mean and the variance of the two chains are the same.

Gelman-Rubin Diagnostic

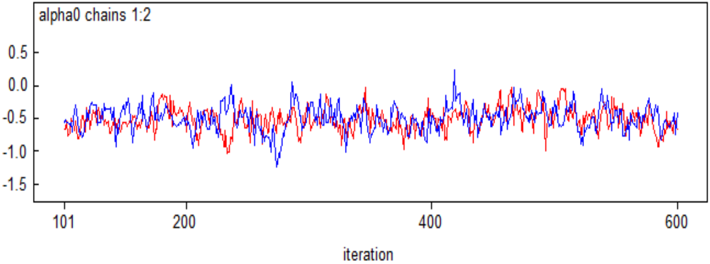
- Gelman (especially) argues that the best way to identify non-convergence is to simulate multiple sequences for over-dispersed starting points.
- The intuition is that the behavior of all of the chains should be basically the same.
- Or, as Gelman and Rubin put it, the variance within the chains should be the same as the variance across the chains.
- I think this can be diagnosed pretty easily through traceplots of multiple chains. You want to see if it looks like the mean and the variance of the two chains are the same.

- Gelman (especially) argues that the best way to identify non-convergence is to simulate multiple sequences for over-dispersed starting points.
- The intuition is that the behavior of all of the chains should be basically the same.
- Or, as Gelman and Rubin put it, the variance within the chains should be the same as the variance across the chains.
- I think this can be diagnosed pretty easily through traceplots of multiple chains. You want to see if it looks like the mean and the variance of the two chains are the same.

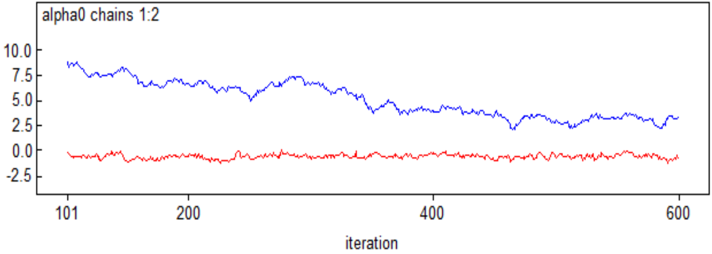
Gelman-Rubin Diagnostic

- Gelman (especially) argues that the best way to identify non-convergence is to simulate multiple sequences for over-dispersed starting points.
- The intuition is that the behavior of all of the chains should be basically the same.
- Or, as Gelman and Rubin put it, the variance within the chains should be the same as the variance across the chains.
- I think this can be diagnosed pretty easily through traceplots of multiple chains. You want to see if it looks like the mean and the variance of the two chains are the same.

Examples where convergence seems reasonable



Examples where convergence seems unreasonable



Gelman-Rubin Diagnostic

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

Gelman-Rubin Diagnostic

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

Gelman-Rubin Diagnostic

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     - 2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     - 2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     - 2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     - 2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

Gelman-Rubin Diagnostic

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     - 2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     - 2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     - 2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     - 2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

- The Gelman-Rubin statistic is based on the following procedure:
  1. estimate your model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.
  2. take the n-monitored draws of m parameters and calculate the following statistics:
     2.1 Within chain variance $W = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2$
     2.2 Between chain variance $B = \frac{m}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$
     2.3 Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$
     2.4 Gelman-Rubin Statistics: $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$

Gelman Rubin Diagnostic

- Before convergence, W underestimates total posterior variance in θ because it has not fully explored the target distribution.
- $V(\theta)$ on the other hand overestimates variance in θ because the starting points are over-dispersed relative to the target.
- Once convergence is reached, W and $V(\theta)$ should be almost equivalent because variation within the chains and variations between the chains should coincide, so R should approximately equal one.
- The drawback of this stat is that its value depends a great deal on the choice of initial values.

Gelman Rubin Diagnostic

- Before convergence, W underestimates total posterior variance in θ because it has not fully explored the target distribution.

- $V(\theta)$ on the other hand overestimates variance in θ because the starting points are over-dispersed relative to the target.

- Once convergence is reached, W and $V(\theta)$ should be almost equivalent because variation within the chains and variations between the chains should coincide, so R should approximately equal one.

- The drawback of this stat is that its value depends a great deal on the choice of initial values.

Gelman Rubin Diagnostic

- Before convergence, W underestimates total posterior variance in θ because it has not fully explored the target distribution.
- $V(\theta)$ on the other hand overestimates variance in θ because the starting points are over-dispersed relative to the target.
- Once convergence is reached, W and $V(\theta)$ should be almost equivalent because variation within the chains and variations between the chains should coincide, so R should approximately equal one.
- The drawback of this stat is that its value depends a great deal on the choice of initial values.

Gelman Rubin Diagnostic

- Before convergence, W underestimates total posterior variance in θ because it has not fully explored the target distribution.
- $V(\theta)$ on the other hand overestimates variance in θ because the starting points are over-dispersed relative to the target.
- Once convergence is reached, W and $V(\theta)$ should be almost equivalent because variation within the chains and variations between the chains should coincide, so R should approximately equal one.
- The drawback of this stat is that its value depends a great deal on the choice of initial values.

Convergence Diagnostic Summary

- You can never prove that something has converged, you can only tell when something has not converged.
- If your model has not converged and you are confident that YOU have not made any mistake, then the best thing is to just let the model run a long time. CPU time is often "cheaper" than your time.
- For models with large numbers of parameters you should let the model run for a long time.
- There are a number of easy way to implement tricks (mostly reparamerizations) that will help to speed convergence in most regression-based models.
- **Convergence does not mean that you have a good model!!! Convergence should be the beginning of model assessment, not its end.**

Convergence Diagnostic Summary

- You can never prove that something has converged, you can only tell when something has not converged.
- If your model has not converged and you are confident that YOU have not made any mistake, then the best thing is to just let the model run a long time. CPU time is often "cheaper" than your time.
- For models with large numbers of parameters you should let the model run for a long time.
- There are a number of easy way to implement tricks (mostly reparamerizations) that will help to speed convergence in most regression-based models.
- **Convergence does not mean that you have a good model!!! Convergence should be the beginning of model assessment, not its end.**

Convergence Diagnostic Summary

- You can never prove that something has converged, you can only tell when something has not converged.
- If your model has not converged and you are confident that YOU have not made any mistake, then the best thing is to just let the model run a long time. CPU time is often "cheaper" than your time.
- For models with large numbers of parameters you should let the model run for a long time.
- There are a number of easy way to implement tricks (mostly reparamerizations) that will help to speed convergence in most regression-based models.
- **Convergence does not mean that you have a good model!!! Convergence should be the beginning of model assessment, not its end.**

Convergence Diagnostic Summary

- You can never prove that something has converged, you can only tell when something has not converged.
- If your model has not converged and you are confident that YOU have not made any mistake, then the best thing is to just let the model run a long time. CPU time is often "cheaper" than your time.
- For models with large numbers of parameters you should let the model run for a long time.
- There are a number of easy way to implement tricks (mostly reparamerizations) that will help to speed convergence in most regression-based models.
- **Convergence does not mean that you have a good model!!! Convergence should be the beginning of model assessment, not its end.**

Convergence Diagnostic Summary

- You can never prove that something has converged, you can only tell when something has not converged.
- If your model has not converged and you are confident that YOU have not made any mistake, then the best thing is to just let the model run a long time. CPU time is often "cheaper" than your time.
- For models with large numbers of parameters you should let the model run for a long time.
- There are a number of easy way to implement tricks (mostly reparamerizations) that will help to speed convergence in most regression-based models.
- **Convergence does not mean that you have a good model!!! Convergence should be the beginning of model assessment, not its end.**

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

MHRD Govt. of India

सत्यमेव जयते

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Hypothesis Testing and Bayes Factors

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Bayesian p-values
2. Bayes Factors for model comparison
3. Easy to implement alternatives for model comparison

Outline

1. Bayesian p-values
2. Bayes Factors for model comparison
3. Easy to implement alternatives for model comparison

Outline

1. Bayesian p-values
2. Bayes Factors for model comparison
3. Easy to implement alternatives for model comparison

Bayesian Hypothesis Testing

- Bayesian hypothesis testing is less formal than frequentist approach.

- In fact, Bayesian researchers typically summarize the posterior distribution without applying a rigid decision process.

- If one wanted to apply a formal process, Bayesian decision theory is the way to go because it is possible to get a probability distribution over the parameter space and one can make expected utility calculations based on the costs and benefits of different outcomes.

- Considerable energy has been given, however, in trying to map Bayesian statistical models into the null hypothesis hypothesis testing framework, with mixed results at best.

- Bayesian hypothesis testing is less formal than frequentist approach.
- In fact, Bayesian researchers typically summarize the posterior distribution without applying a rigid decision process.
- If one wanted to apply a formal process, Bayesian decision theory is the way to go because it is possible to get a probability distribution over the parameter space and one can make expected utility calculations based on the costs and benefits of different outcomes.
- Considerable energy has been given, however, in trying to map Bayesian statistical models into the null hypothesis hypothesis testing framework, with mixed results at best.

Bayesian Hypothesis Testing

- Bayesian hypothesis testing is less formal than frequentist approach.
- In fact, Bayesian researchers typically summarize the posterior distribution without applying a rigid decision process.
- If one wanted to apply a formal process, Bayesian decision theory is the way to go because it is possible to get a probability distribution over the parameter space and one can make expected utility calculations based on the costs and benefits of different outcomes.
- Considerable energy has been given, however, in trying to map Bayesian statistical models into the null hypothesis hypothesis testing framework, with mixed results at best.

- Bayesian hypothesis testing is less formal than frequentist approach.
- In fact, Bayesian researchers typically summarize the posterior distribution without applying a rigid decision process.
- If one wanted to apply a formal process, Bayesian decision theory is the way to go because it is possible to get a probability distribution over the parameter space and one can make expected utility calculations based on the costs and benefits of different outcomes.
- Considerable energy has been given, however, in trying to map Bayesian statistical models into the null hypothesis hypothesis testing framework, with mixed results at best.

Similarities between Bayesian and Frequentist Hypothesis Testing

- Maximum likelihood estimates of parameter means and standard errors and Bayesian estimates with flat priors are equivalent.
- Asymptotically, the data will overwhelm the choice of prior, so if we had infinite data sets, priors would be irrelevant and Bayesian and frequentist results would converge.
- Frequentist one-tailed tests are basically equivalent to what a Bayesian would get using credible intervals.

Similarities between Bayesian and Frequentist Hypothesis Testing

- Maximum likelihood estimates of parameter means and standard errors and Bayesian estimates with flat priors are equivalent.

- Asymptotically, the data will overwhelm the choice of prior, so if we had infinite data sets, priors would be irrelevant and Bayesian and frequentist results would converge.

- Frequentist one-tailed tests are basically equivalent to what a Bayesian would get using credible intervals.

Similarities between Bayesian and Frequentist Hypothesis Testing

- Maximum likelihood estimates of parameter means and standard errors and Bayesian estimates with flat priors are equivalent.

- Asymptotically, the data will overwhelm the choice of prior, so if we had infinite data sets, priors would be irrelevant and Bayesian and frequentist results would converge.

- Frequentist one-tailed tests are basically equivalent to what a Bayesian would get using credible intervals.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.
- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero.
  e.g., two-sided p-value$=P(-e < B < e)$
- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.
- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.

- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero. e.g., two-sided p-value=$P(-e < B < e)$

- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.

- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.

- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero.
  e.g., two-sided p-value$=P(-e < B < e)$

- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.

- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.

- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero.
  e.g., two-sided p-value$=P(-e < B < e)$

- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.

- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.

- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero.
  e.g., two-sided p-value$=P(-e < B < e)$

- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.

- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

Differences between Frequentist and Bayesian Hypothesis Testing

- The most important pragmatic difference between Bayesian and frequentist hypothesis testing is that Bayesian methods are poorly suited for two-tailed tests.

- Why? Because the probability of zero in a continuous distribution is zero.
  The best solution proposed so far is to calculate the probability that, say, a regression coefficient is in some range near zero.
  e.g., two-sided p-value$=P(-e < B < e)$

- However, the choice of e seems very *ad hoc* unless there is some decision theoretic basis.

- The other important difference is more philosophical. Frequentist p-values violate the likelihood principle.

- Bayes Factors are the dominant method of Bayesian model testing. They are the Bayesian analogues of likelihood ratio tests.

- The basic intuition is that prior and posterior information are combined in a ratio that provides evidence in favor of one model specification verses another.

- Bayes Factors are very flexible, allowing multiple hypotheses to be compared simultaneously and nested models are not required in order to make comparisons $\longrightarrow$ it goes without saying that compared models should obviously have the same dependent variable.

- Bayes Factors are the dominant method of Bayesian model testing. They are the Bayesian analogues of likelihood ratio tests.

- The basic intuition is that prior and posterior information are combined in a ratio that provides evidence in favor of one model specification verses another.

- Bayes Factors are very flexible, allowing multiple hypotheses to be compared simultaneously and nested models are not required in order to make comparisons $\longrightarrow$ it goes without saying that compared models should obviously have the same dependent variable.

- Bayes Factors are the dominant method of Bayesian model testing. They are the Bayesian analogues of likelihood ratio tests.

- The basic intuition is that prior and posterior information are combined in a ratio that provides evidence in favor of one model specification verses another.

- Bayes Factors are very flexible, allowing multiple hypotheses to be compared simultaneously and nested models are not required in order to make comparisons $\longrightarrow$ it goes without saying that compared models should obviously have the same dependent variable.

The General Form for Bayes Factors

- Suppose that we observe data X and with to test two competing models - M_1 and M_2 relating these data to two different sets of parameters θ_1 and θ_2.

- We would like to know which of the following likelihood specifications is better:

$$M_1: \ f_1(x|\theta_1) \text{ and } M_2: \ f_2(x|\theta_2)$$

- Obviously, we would need prior distributions for the θ_1 and θ_2 and prior probabilities for M_1 and M_2

The General Form for Bayes Factors

- Suppose that we observe data X and with to test two competing models - M_1 and M_2 relating these data to two different sets of parameters θ_1 and θ_2.

- We would like to know which of the following likelihood specifications is better:

$$M_1: \ f_1(x|\theta_1) \text{ and } M_2: \ f_2(x|\theta_2)$$

- Obviously, we would need prior distributions for the θ_1 and θ_2 and prior probabilities for M_1 and M_2

The General Form for Bayes Factors

- Suppose that we observe data X and with to test two competing models - M_1 and M_2 relating these data to two different sets of parameters θ_1 and θ_2.

- We would like to know which of the following likelihood specifications is better:

$$M_1 : \; f_1(x|\theta_1) \text{ and } M_2 : \; f_2(x|\theta_2)$$

- Obviously, we would need prior distributions for the θ_1 and θ_2 and prior probabilities for M_1 and M_2

The General Form for Bayes Factors

- The posterior odds ratio in favor of M_1 over M_2 is:

  Posterior Odds=Prior Odds/Data $\times$ Bayes factor

$$\frac{\pi(M_1|x)}{\pi(M_2|x)} = \frac{p(M_1)/p(x)}{p(M_2)/p(x)} \times \frac{\int_{\theta_1} f_1(x|\theta_1)p_1(\theta_1)d\theta_1}{\int_{\theta_2} f_2(x|\theta_2)p_2(\theta_2)d\theta_2}$$

- Rearranging terms, we find the Bayes' factor is:

$$\text{Bayes Factor} = B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- If we have nested models and $P(M_1) = P(M_2) = 0.5$ then the Bayes factor reduces to likelihood ratio

The General Form for Bayes Factors

- The posterior odds ratio in favor of M_1 over M_2 is:

  Posterior Odds=Prior Odds/Data $\times$ Bayes factor

$$\frac{\pi(M_1|x)}{\pi(M_2|x)} = \frac{p(M_1)/p(x)}{p(M_2)/p(x)} \times \frac{\int_{\theta_1} f_1(x|\theta_1)p_1(\theta_1)d\theta_1}{\int_{\theta_2} f_2(x|\theta_2)p_2(\theta_2)d\theta_2}$$

- Rearranging terms, we find the Bayes' factor is:

$$\text{Bayes Factor} = B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- If we have nested models and $P(M_1) = P(M_2) = 0.5$ then the Bayes factor reduces to likelihood ratio

The General Form for Bayes Factors

- The posterior odds ratio in favor of M_1 over M_2 is:

  Posterior Odds=Prior Odds/Data $\times$ Bayes factor

$$\frac{\pi(M_1|x)}{\pi(M_2|x)} = \frac{p(M_1)/p(x)}{p(M_2)/p(x)} \times \frac{\int_{\theta_1} f_1(x|\theta_1)p_1(\theta_1)d\theta_1}{\int_{\theta_2} f_2(x|\theta_2)p_2(\theta_2)d\theta_2}$$

- Rearranging terms, we find the Bayes' factor is:

$$\text{Bayes Factor} = B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- If we have nested models and $P(M_1) = P(M_2) = 0.5$ then the Bayes factor reduces to likelihood ratio

- The posterior odds ratio in favor of M_1 over M_2 is:

  Posterior Odds=Prior Odds/Data × Bayes factor

$$\frac{\pi(M_1|x)}{\pi(M_2|x)} = \frac{p(M_1)/p(x)}{p(M_2)/p(x)} \times \frac{\int_{\theta_1} f_1(x|\theta_1)p_1(\theta_1)d\theta_1}{\int_{\theta_2} f_2(x|\theta_2)p_2(\theta_2)d\theta_2}$$

- Rearranging terms, we find the Bayes' factor is:

$$\text{Bayes Factor} = B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- If we have nested models and $P(M_1) = P(M_2) = 0.5$ then the Bayes factor reduces to likelihood ratio

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

Rule of Thumb

- Bayes Factor :

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)}$$

- With this setup, if we interpret model 1 as the null model, then:
  1. If $B(x) \geq 1$ then model 1 is supported
  2. If $1 > B(x) \geq 10^{-1/2}$ then minimal evidence against model 1.
  3. If $10^{-1/2} > B(x) \geq 10^{-1}$ then substantial evidence against model 1.
  4. If $10^{-1} > B(x) \geq 10^{-2}$ then strong evidence against model 1.
  5. If $10^{-2} > B(x)$ then decisive evidence against model 1.

The Bad News

- Unfortunately, while Bayes Factors are rather intuitive, as a practical matter they are often quite difficult to calculate.

- However, in the MCMCpack package Bayes Factor can be computed routinely for standard statistical models

- You also may want to use Carlin and Chib's technique for computing Bayes Factors for competing non-nested regression models reported in *Journal of Royal Statistical Society. Series B.* vol 57:3 1995.

- Our discussion will focus on alternatives to the Bayes Factor.

- Unfortunately, while Bayes Factors are rather intuitive, as a practical matter they are often quite difficult to calculate.
- However, in the MCMCpack package Bayes Factor can be computed routinely for standard statistical models
- You also may want to use Carlin and Chib's technique for computing Bayes Factors for competing non-nested regression models reported in *Journal of Royal Statistical Society. Series B.* vol 57:3 1995.
- Our discussion will focus on alternatives to the Bayes Factor.

The Bad News

- Unfortunately, while Bayes Factors are rather intuitive, as a practical matter they are often quite difficult to calculate.
- However, in the MCMCpack package Bayes Factor can be computed routinely for standard statistical models
- You also may want to use Carlin and Chib's technique for computing Bayes Factors for competing non-nested regression models reported in *Journal of Royal Statistical Society. Series B.* vol 57:3 1995.
- Our discussion will focus on alternatives to the Bayes Factor.

The Bad News

- Unfortunately, while Bayes Factors are rather intuitive, as a practical matter they are often quite difficult to calculate.
- However, in the MCMCpack package Bayes Factor can be computed routinely for standard statistical models
- You also may want to use Carlin and Chib's technique for computing Bayes Factors for competing non-nested regression models reported in *Journal of Royal Statistical Society. Series B.* vol 57:3 1995.
- Our discussion will focus on alternatives to the Bayes Factor.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- **Good**: The Likelihood Ratio

$$Ratio = -2[\log\ L(\theta^*_{Restricted\ Model}|y) - \log\ L(\theta^*_{Full\ Model}|y)]$$

- This statistic will always favor the unrestricted model, but when the Bayes estimators or equivalent to the maximum likelihood estimates, then the Ratio is distributed as a χ^2 where the number of degrees of freedom is equal to the number of test parameters.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- **Good**: The Likelihood Ratio

$$Ratio = -2[\log\ L(\theta^*_{Restricted\ Model}|y) - \log\ L(\theta^*_{Full\ Model}|y)]$$

- This statistic will always favor the unrestricted model, but when the Bayes estimators or equivalent to the maximum likelihood estimates, then the Ratio is distributed as a χ^2 where the number of degrees of freedom is equal to the number of test parameters.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- **Good**: The Likelihood Ratio

$$Ratio = -2[\log\ L(\theta^*_{Restricted\ Model}|y) - \log\ L(\theta^*_{Full\ Model}|y)]$$

- This statistic will always favor the unrestricted model, but when the Bayes estimators or equivalent to the maximum likelihood estimates, then the Ratio is distributed as a χ^2 where the number of degrees of freedom is equal to the number of test parameters.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- Better: Akaike Information Criterion (AIC)
  $AIC = -2\log\ L(\theta^*|y) + 2p$

- where $p =$ the number of parameters including the intercept.

- To compare two models, compare the AIC from model 1 against the AIC from model 2.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- Better: Akaike Information Criterion (AIC)
  $AIC = -2\log\ L(\theta^*|y) + 2p$

- where $p =$ the number of parameters including the intercept.

- To compare two models, compare the AIC from model 1 against the AIC from model 2.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- Better: Akaike Information Criterion (AIC)
  $AIC = -2\log\ L(\theta^*|y) + 2p$

- where $p =$ the number of parameters including the intercept.

- To compare two models, compare the AIC from model 1 against the AIC from model 2.

Alternatives to the Bayes Factor for model assessmen

- Let θ^* denote your estimates of the parameter means (or medians or modes) in your model and suppose that the Bayes estimate is approximately equal to the maximum likelihood estimate, then the following stats used in frequentist statistics will be useful diagnostics.

- Better: Akaike Information Criterion (AIC)
  $AIC = -2\log\ L(\theta^*|y) + 2p$

- where $p =$ the number of parameters including the intercept.

- To compare two models, compare the AIC from model 1 against the AIC from model 2.

Alternatives to the Bayes Factor for model assessmen

- Models do not need to be nested
- The AIC tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.
- Bayesian Information Criterion (BIC):

$$BIC = -2\log\ L(\theta^*|y) + p \times \log(n)$$

where p is the number of parameters and n is the sample size.

- This statistic can also be used for non-nested models
- $BIC_1 - BIC_2 \approx -2log(Bayes\ Factor_{12})$ for Model 1 vs Model 2

Alternatives to the Bayes Factor for model assessmen

- Models do not need to be nested
- The AIC tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.
- Bayesian Information Criterion (BIC):

$$BIC = -2\log\ L(\theta^*|y) + p \times \log(n)$$

where p is the number of parameters and n is the sample size.

- This statistic can also be used for non-nested models
- $BIC_1 - BIC_2 \approx -2log(Bayes\ Factor_{12})$ for Model 1 vs Model 2

Alternatives to the Bayes Factor for model assessmen

- Models do not need to be nested
- The AIC tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.
- Bayesian Information Criterion (BIC):

$$BIC = -2\log\ L(\theta^*|y) + p \times \log(n)$$

where p is the number of parameters and n is the sample size.

- This statistic can also be used for non-nested models
- $BIC_1 - BIC_2 \approx -2log(Bayes\ Factor_{12})$ for Model 1 vs Model 2

Alternatives to the Bayes Factor for model assessmen

- Models do not need to be nested

- The AIC tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.

- Bayesian Information Criterion (BIC):

$$BIC = -2\log\ L(\theta^*|y) + p \times \log(n)$$

where p is the number of parameters and n is the sample size.

- This statistic can also be used for non-nested models

- $BIC_1 - BIC_2 \approx -2log(Bayes\ Factor_{12})$ for Model 1 vs Model 2

Alternatives to the Bayes Factor for model assessmen

- Models do not need to be nested

- The AIC tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.

- Bayesian Information Criterion (BIC):

$$BIC = -2\log\ L(\theta^*|y) + p \times \log(n)$$

  where p is the number of parameters and n is the sample size.

- This statistic can also be used for non-nested models

- $BIC_1 - BIC_2 \approx -2log(Bayes\ Factor_{12})$ for Model 1 vs Model 2

MHRD Govt. of India

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$\begin{aligned} z_i = \beta_0 & + \beta_1 Age_i + \beta_2 I(race_i = black) \\ & + \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$z_i = \beta_0 + \beta_1 Age_i + \beta_2 I(race_i = black) + \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

- Consider 'birthwt' dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2..., 189$

$$\begin{aligned} z_i = \beta_0 &+ \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

- Consider ‘birthwt’ dataset available in MASS package of R
- The dataset tries to look for the risk factors associated with low infant birth weight.

$$low_i = \begin{cases} 1 & \text{indicator of birth weight less than 2.5 kg.} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, 2..., 189$$

$$\begin{aligned} z_i = \beta_0 &+ \beta_1 Age_i + \beta_2 I(race_i = black) \\ &+ \beta_3 I(race_i = others) + \beta_4 I(Smoke_i = yes) + \epsilon_i \end{aligned}$$

$\epsilon_i \sim N(0, 1)$ and
$P(low = 1|Age, race, smoke) = P(z > 0|Age, race, smoke)$

Posterior Summary

```
> library(MCMCpack)
> set.seed(8135)
> data(birthwt)
> M1 <- MCMCprobit(low~as.factor(race)+age+smoke
+                 , data=birthwt, b0 = 0, B0 = 10
+                 ,marginal.likelihood="Chib95")
> M2 <- MCMCprobit(low~as.factor(race) +smoke
+                 , data=birthwt, b0 = 0, B0 = 10
+                 ,marginal.likelihood="Chib95")
> M3 <- MCMCprobit(low~as.factor(race) +age
+                 , data=birthwt, b0 = 0 , B0 = 10
+                 ,marginal.likelihood="Chib95")
```

Posterior Summary

```
> BF <- BayesFactor(M1, M2, M3)
> round(BF$BF.mat,digit=3)
```

	M1	M2	M3
M1	1.000	1.445	6.807
M2	0.692	1.000	4.711
M3	0.147	0.212	1.000

- $BF_{1,2} = 1.445$ indicates that data occurred 1.41 times more likely under Model 1 (M1) than Model 2 (M2). It can be considered as an anecdotal evidence
- $BF_{1,3} = 6.807$ indicates that data occurred 6.81 times more likely under Model 1 (M1) than Model 3 (M3). It can be considered as moderate evidence
- $BF_{2,3} = 4.711$ indicates that data occurred 4.71 times more likely under Model 2 (M2) than Model 3 (M3).

Posterior Summary

```
> BF <- BayesFactor(M1, M2, M3)
> round(BF$BF.mat,digit=3)
```

	M1	M2	M3
M1	1.000	1.445	6.807
M2	0.692	1.000	4.711
M3	0.147	0.212	1.000

- $BF_{1,2} = 1.445$ indicates that data occurred 1.41 times more likely under Model 1 (M1) than Model 2 (M2). It can be considered as an anecdotal evidence
- $BF_{1,3} = 6.807$ indicates that data occurred 6.81 times more likely under Model 1 (M1) than Model 3 (M3). It can be considered as moderate evidence
- $BF_{2,3} = 4.711$ indicates that data occurred 4.71 times more likely under Model 2 (M2) than Model 3 (M3).

Posterior Summary

```
> BF <- BayesFactor(M1, M2, M3)
> round(BF$BF.mat,digit=3)
```

	M1	M2	M3
M1	1.000	1.445	6.807
M2	0.692	1.000	4.711
M3	0.147	0.212	1.000

- $BF_{1,2} = 1.445$ indicates that data occurred 1.41 times more likely under Model 1 (M1) than Model 2 (M2). It can be considered as an anecdotal evidence
- $BF_{1,3} = 6.807$ indicates that data occurred 6.81 times more likely under Model 1 (M1) than Model 3 (M3). It can be considered as moderate evidence
- $BF_{2,3} = 4.711$ indicates that data occurred 4.71 times more likely under Model 2 (M2) than Model 3 (M3).

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Advanced Hierarchical Models - Part 1

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

1. The intuition behind hierarchical regression models
2. Setting up probability models for hierarchical regressions

Outline

1. The intuition behind hierarchical regression models

2. Setting up probability models for hierarchical regressions

Overview

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
- e.g. we collect measurements by geographic region or social group.
- Hierarchical models provide a way of examining differences across populations. They pool the information for the disparate groups without assuming that they belong to precisely the same population.
- In the context of regression analyses, hierarchical models allow us to examine whether the extent to which regression coefficients vary across different sub-populations, while borrowing strength from the full sample.

Overview

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
- e.g. we collect measurements by geographic region or social group.
- Hierarchical models provide a way of examining differences across populations. They pool the information for the disparate groups without assuming that they belong to precisely the same population.
- In the context of regression analyses, hierarchical models allow us to examine whether the extent to which regression coefficients vary across different sub-populations, while borrowing strength from the full sample.

Overview

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
- e.g. we collect measurements by geographic region or social group.
- Hierarchical models provide a way of examining differences across populations. They pool the information for the disparate groups without assuming that they belong to precisely the same population.
- In the context of regression analyses, hierarchical models allow us to examine whether the extent to which regression coefficients vary across different sub-populations, while borrowing strength from the full sample.

Overview

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
- e.g. we collect measurements by geographic region or social group.
- Hierarchical models provide a way of examining differences across populations. They pool the information for the disparate groups without assuming that they belong to precisely the same population.
- In the context of regression analyses, hierarchical models allow us to examine whether the extent to which regression coefficients vary across different sub-populations, while borrowing strength from the full sample.

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- Key modeling question: Does the sample pool?

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).
- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.
- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.
  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.
- Key modeling question: Does the sample pool?

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.
  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- Key modeling question: Does the sample pool?

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.
  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.
- Key modeling question: Does the sample pool?

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.
  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- Key modeling question: Does the sample pool?

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).
- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.
- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.
  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.
- Key modeling question: Does the sample pool?

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.
- Key modeling question: Does the sample pool?

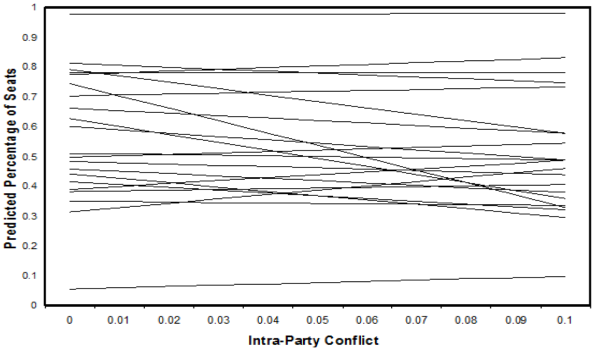
Parameters of Pooled OLS Model of Democratic Electoral Success Due to Intra-Party Unity

Dep. Var. Democratic Electoral Success	Posteriod Mean	Posterior standard deviation
Intercept	0.578^{*}	0.024
Ideological Conflict	-3.512^{*}	1.192

- Mean Squared Error: 0.08516

  * Denotes statistical significance

Unpooled OLS Model
(Different state-specific intercepts and slop



- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model. What to do?

Example

- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model.

- In a context like this, hierarchical structures are perfect.
  - Where differences are not statistically important, the state-specific coefficients are shrunk back toward the national average.

  - Where differences are statistically meaningful, the state-specific effects remain markedly different from the national average.

Example

- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model.

- In a context like this, hierarchical structures are perfect.
  - Where differences are not statistically important, the state-specific coefficients are shrunk back toward the national average.
  - Where differences are statistically meaningful, the state-specific effects remain markedly different from the national average.

Example

- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model.

- In a context like this, hierarchical structures are perfect.
  - Where differences are not statistically important, the state-specific coefficients are shrunk back toward the national average.
  - Where differences are statistically meaningful, the state-specific effects remain markedly different from the national average.

Example

- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model.

- In a context like this, hierarchical structures are perfect.
  - Where differences are not statistically important, the state-specific coefficients are shrunk back toward the national average.

  - Where differences are statistically meaningful, the state-specific effects remain markedly different from the national average.

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i * \textit{Intra-Party Conflict}_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i * \textit{Intra-Party Conflict}_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i *$ *Intra-Party Conflict*$_{i,t-1}$,
- $a_i \sim Norna(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim Gamma(0.1, 0.1)$
- $\sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim Gamma(0.1, 0.1)$
- and $\tau \sim Gamma(0.1, 0.1)$

- The crucial difference between unpooled OLS and the hierarchical model is that the state-specific intercept terms and the coefficients for intra-party conflict are now treated as exchangeable draws from a common probability model with unknown mean and variance.

- The posterior distributions of these state-specific parameters convey information about local effects.

- The hyper-parameter A represents the average level of Democratic electoral success while τ_A measures the variation in the party's fortunes across states.

- Similarly, B is the average impact of intra-party conflict, while τ_B indicates the variation in the influence of party unity across states.

Comments

- The crucial difference between unpooled OLS and the hierarchical model is that the state-specific intercept terms and the coefficients for intra-party conflict are now treated as exchangeable draws from a common probability model with unknown mean and variance.

- The posterior distributions of these state-specific parameters convey information about local effects.

- The hyper-parameter A represents the average level of Democratic electoral success while τ_A measures the variation in the party's fortunes across states.

- Similarly, B is the average impact of intra-party conflict, while τ_B indicates the variation in the influence of party unity across states.

Comments

- The crucial difference between unpooled OLS and the hierarchical model is that the state-specific intercept terms and the coefficients for intra-party conflict are now treated as exchangeable draws from a common probability model with unknown mean and variance.

- The posterior distributions of these state-specific parameters convey information about local effects.

- The hyper-parameter A represents the average level of Democratic electoral success while τ_A measures the variation in the party's fortunes across states.

- Similarly, B is the average impact of intra-party conflict, while τ_B indicates the variation in the influence of party unity across states.

Comments

- The crucial difference between unpooled OLS and the hierarchical model is that the state-specific intercept terms and the coefficients for intra-party conflict are now treated as exchangeable draws from a common probability model with unknown mean and variance.

- The posterior distributions of these state-specific parameters convey information about local effects.

- The hyper-parameter A represents the average level of Democratic electoral success while τ_A measures the variation in the party's fortunes across states.

- Similarly, B is the average impact of intra-party conflict, while τ_B indicates the variation in the influence of party unity across states.

Comments

- If the posterior distribution of the hyper-parameters reveal that $\tau_A = \tau_B = \infty$ then pooled OLS is a special case.

- This is because if there is no variance (i.e. infinite precision) in the intercept or coefficient across states, then one should conclude that there are no regime effects.

- Similarly, if $\tau_A = \tau_B = 0$, then unpooled OLS is a special case because there is no underlying structure to the data across states.

Comments

- If the posterior distribution of the hyper-parameters reveal that $\tau_A = \tau_B = \infty$ then pooled OLS is a special case.
- This is because if there is no variance (i.e. infinite precision) in the intercept or coefficient across states, then one should conclude that there are no regime effects.
- Similarly, if $\tau_A = \tau_B = 0$, then unpooled OLS is a special case because there is no underlying structure to the data across states.

Comments

- If the posterior distribution of the hyper-parameters reveal that $\tau_A = \tau_B = \infty$ then pooled OLS is a special case.

- This is because if there is no variance (i.e. infinite precision) in the intercept or coefficient across states, then one should conclude that there are no regime effects.

- Similarly, if $\tau_A = \tau_B = 0$, then unpooled OLS is a special case because there is no underlying structure to the data across states.

Hyper-Parameters for Model of Democratic Electoral Success Due to Intra-Party Unity

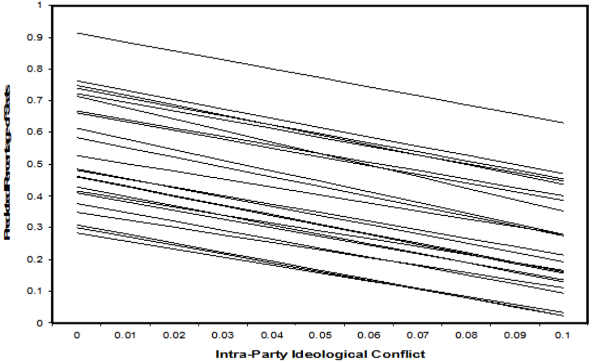
Parameter for State specific Intercept	Posteriod Mean	Posterior standard deviation
Mean of Stat specific Intercept	0.54^*	0.054
Precision of State Specific Intercept	21.2	8.101

* denotes statistical significance

Parameter for State specific Intra-Part Coefficients	Posteriod Mean	Posterior standard deviation
Mean of Stat specific Coefficients	-2.85^*	1.216
Precision of State Specific Coefficients	3.071	4.81

* denotes statistical significance

State-Specific Predicted Values



What sort of voodoo is this?

- The explanation for why the random coefficient model had such a substantial impact on the parameter estimates for intra-party conflict was precisely because our pooling tests rejected the joint significance of state-specific effects.
- The wild variations observed from unpooled OLS were an artifact of over-fitting the data based on a small number of observations.
- To prevent this over-fitting, the random coefficient model "borrowed strength" from the overall effect of the independent variable in order to make inferences about the state-specific effects.

What sort of voodoo is this?

- The explanation for why the random coefficient model had such a substantial impact on the parameter estimates for intra-party conflict was precisely because our pooling tests rejected the joint significance of state-specific effects.
- The wild variations observed from unpooled OLS were an artifact of over-fitting the data based on a small number of observations.
- To prevent this over-fitting, the random coefficient model "borrowed strength" from the overall effect of the independent variable in order to make inferences about the state-specific effects.

What sort of voodoo is this?

- The explanation for why the random coefficient model had such a substantial impact on the parameter estimates for intra-party conflict was precisely because our pooling tests rejected the joint significance of state-specific effects.
- The wild variations observed from unpooled OLS were an artifact of over-fitting the data based on a small number of observations.
- To prevent this over-fitting, the random coefficient model "borrowed strength" from the overall effect of the independent variable in order to make inferences about the state-specific effects.

What sort of voodoo is this?

- The extent of this borrowing is contingent on the relative precision of the state-specific and overall effects.

- Thus, the regression lines became approximately parallel with the introduction of the random coefficient model because there was relatively little information provided by the state-specific data regarding the effect of intra-party conflict relative to that provided by the entire sample.

- Meanwhile, the intercepts remained variant across regression lines, because there was sufficient state-specific data to establish that each state had a different predisposition in favor or against the Democratic Party.

What sort of voodoo is this?

- The extent of this borrowing is contingent on the relative precision of the state-specific and overall effects.
- Thus, the regression lines became approximately parallel with the introduction of the random coefficient model because there was relatively little information provided by the state-specific data regarding the effect of intra-party conflict relative to that provided by the entire sample.
- Meanwhile, the intercepts remained variant across regression lines, because there was sufficient state-specific data to establish that each state had a different predisposition in favor or against the Democratic Party.

What sort of voodoo is this?

- The extent of this borrowing is contingent on the relative precision of the state-specific and overall effects.
- Thus, the regression lines became approximately parallel with the introduction of the random coefficient model because there was relatively little information provided by the state-specific data regarding the effect of intra-party conflict relative to that provided by the entire sample.
- Meanwhile, the intercepts remained variant across regression lines, because there was sufficient state-specific data to establish that each state had a different predisposition in favor or against the Democratic Party.

Example:

- conside the data set called cheese from the baysem package.

- The data set contains marketing data of certain brand name processed cheese, such as the weekly sales volume (VOLUME), unit retail price (PRICE), and display activity level (DISP) in various regional retailer accounts.

Example:

- conside the data set called cheese from the baysem package.
- The data set contains marketing data of certain brand name processed cheese, such as the weekly sales volume (VOLUME), unit retail price (PRICE), and display activity level (DISP) in various regional retailer accounts.

Example:

- For each account, we can define the following linear regression model of the log sales volume, where β_1 is the intercept term, β_2 is the display measure coefficient, and β_3 is the log price coefficient.

$$\log(Volume) = \beta_1 + \beta_2 * Display + \beta_3 * \log(Price) + \epsilon$$

ϵ relies on regional market conditions, and we would not expect it to have the same dispersion among retailers.

- For the same reason, we cannot expect identical regression coefficients for all accounts, or attempt to define a single linear regression model for the entire data set.

Example:

- For each account, we can define the following linear regression model of the log sales volume, where β_1 is the intercept term, β_2 is the display measure coefficient, and β_3 is the log price coefficient.

$$\log(Volume) = \beta_1 + \beta_2 * Display + \beta_3 * \log(Price) + \epsilon$$

ϵ relies on regional market conditions, and we would not expect it to have the same dispersion among retailers.

- For the same reason, we cannot expect identical regression coefficients for all accounts, or attempt to define a single linear regression model for the entire data set.

- For each account, we can define the following linear regression model of the log sales volume, where β_1 is the intercept term, β_2 is the display measure coefficient, and β_3 is the log price coefficient.

$$\log(Volume) = \beta_1 + \beta_2 * Display + \beta_3 * \log(Price) + \epsilon$$

ϵ relies on regional market conditions, and we would not expect it to have the same dispersion among retailers.

- For the same reason, we cannot expect identical regression coefficients for all accounts, or attempt to define a single linear regression model for the entire data set.

Example:

- we do expect regression coefficients of the retailer accounts to be related. A common approach to simulate the relationship is the hierarchical linear model, which treats the regression coefficients as random variables of yet another linear regression at the system level.

- **Problem**:

  Fit the data set cheese with the hierarchical linear model, and estimate the average impact on sales volumes of the retailers if the unit retail price is to be raised by 5%

Example:

- we do expect regression coefficients of the retailer accounts to be related. A common approach to simulate the relationship is the hierarchical linear model, which treats the regression coefficients as random variables of yet another linear regression at the system level.

- **Problem**:

  Fit the data set cheese with the hierarchical linear model, and estimate the average impact on sales volumes of the retailers if the unit retail price is to be raised by 5%

Example:

- Let i to be an integer between 1 and the number of retailer accounts. We define a filter for the i^{th} account as follows.

```
> library(bayesm)
> data(cheese)
> retailer<-levels(cheese$RETAILER)
> nreg<-length(retailer)
```

- We now loop through the accounts, and create a list of data items consisting of the X and y components of the linear regression model in each account. The columns of X below contains the intercept placeholder, the display measure, and log price data.

```
> regdata<-NULL
> for (i in 1:nreg) {
+     filter <- cheese$RETAILER==retailer[i]
+     y <- log(cheese$VOLUME[filter])
+     X <- cbind(1,       # intercept placeholder
+             cheese$DISP[filter],
+             log(cheese$PRICE[filter]))
+     regdata[[i]] <- list(y=y, X=X)
+ }
```

- We wrap the regdata and the iteration parameters in lists, and invoke the rhierLinearModel method of the bayesm package. It takes about half a minute for 2,000 MCMC iterations on an average CPU.

```
> Data <- list(regdata=regdata)
> Mcmc <- list(R=2000)
> set.seed(7831)
> system.time(out <- bayesm::rhierLinearModel(
+             Data=Data,
+             Mcmc=Mcmc))
Z not specified -- putting in iota

Starting Gibbs Sampler for Linear Hierarchical Model
    88  Regressions
    1  Variables in Z (if 1, then only intercept)

Prior Parms:
```

Example:

- We can perform the same MCMC simulation with an identically named method in rpudplus. Note the extra output option that we explicitly set to be bayesm for compatibility.

```
> beta.3 <- mean(as.vector(out$betadraw[, 3, 201:2000])
> beta.3
[1] -2.145307
```

- A 5% increase of the unit price amounts to an increment of $\log(1.05)$ in the log price.

- The computation below shows that the sales volume is expected to decrease by 10% on average.

```
> exp(beta.3 * log(1.05))
[1] 0.9006218
```

Example:

- We can perform the same MCMC simulation with an identically named method in rpudplus. Note the extra output option that we explicitly set to be bayesm for compatibility.

  ```
  > beta.3 <- mean(as.vector(out$betadraw[, 3, 201:2000]))
  > beta.3
  ```

  [1] -2.145307

- A 5% increase of the unit price amounts to an increment of $\log(1.05)$ in the log price.

- The computation below shows that the sales volume is expected to decrease by 10% on average.

  ```
  > exp(beta.3 * log(1.05))
  ```

  [1] 0.9006218

Example:

- We can perform the same MCMC simulation with an identically named method in rpudplus. Note the extra output option that we explicitly set to be bayesm for compatibility.

```
> beta.3 <- mean(as.vector(out$betadraw[, 3, 201:2000])
> beta.3
[1] -2.145307
```

- A 5% increase of the unit price amounts to an increment of $\log(1.05)$ in the log price.

- The computation below shows that the sales volume is expected to decrease by 10% on average.

```
> exp(beta.3 * log(1.05))
[1] 0.9006218
```

Example:

- We can perform the same MCMC simulation with an identically named method in rpudplus. Note the extra output option that we explicitly set to be bayesm for compatibility.

```
> beta.3 <- mean(as.vector(out$betadraw[, 3, 201:2000])
> beta.3
[1] -2.145307
```

- A 5% increase of the unit price amounts to an increment of $\log(1.05)$ in the log price.

- The computation below shows that the sales volume is expected to decrease by 10% on average.

```
> exp(beta.3 * log(1.05))
[1] 0.9006218
```

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Advanced Hierarchical Models - Part 2

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Review of the basic hierarchical regression probability model
2. Implementation of Bayesian hierarchical models

1. Review of the basic hierarchical regression probability model
2. Implementation of Bayesian hierarchical models

- Suppose we have a standard multiple regression model where observations i cluster across sub-populations j.
  where j indexes, for example, geographic location, social group, period in history.

- But, we do not want to assume that regression coefficients are identical across sub-populations.

- We also want to allow for unequal variances across sub-populations.

- We assume that each observation is distributed normally with an expected value determined by both observation-specific and sub-population characteristics and level of aggregation-specific variance. Thus,

$$y_{ij} \sim N(m_{ij}, t_j)$$

- Suppose we have a standard multiple regression model where observations i cluster across sub-populations j.
  where j indexes, for example, geographic location, social group, period in history.

- But, we do not want to assume that regression coefficients are identical across sub-populations.

- We also want to allow for unequal variances across sub-populations.

- We assume that each observation is distributed normally with an expected value determined by both observation-specific and sub-population characteristics and level of aggregation-specific variance. Thus,

$$y_{ij} \sim N(m_{ij}, t_j)$$

Hierarchical Regression Models

- Suppose we have a standard multiple regression model where observations i cluster across sub-populations j.
  where j indexes, for example, geographic location, social group, period in history.

- But, we do not want to assume that regression coefficients are identical across sub-populations.

- We also want to allow for unequal variances across sub-populations.

- We assume that each observation is distributed normally with an expected value determined by both observation-specific and sub-population characteristics and level of aggregation-specific variance. Thus,

$$y_{ij} \sim N(m_{ij}, t_j)$$

- Suppose we have a standard multiple regression model where observations i cluster across sub-populations j.
  where j indexes, for example, geographic location, social group, period in history.

- But, we do not want to assume that regression coefficients are identical across sub-populations.

- We also want to allow for unequal variances across sub-populations.

- We assume that each observation is distributed normally with an expected value determined by both observation-specific and sub-population characteristics and level of aggregation-specific variance. Thus,

$$y_{ij} \sim N(m_{ij}, t_j)$$

The Random Coefficient Model

- Suppose

$$y_{ij} \sim N(m_{ij}, t_j)$$

then

$$m_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + ... + \beta_{kj} X_{kj}$$

- For a **random coefficient** (hierarchical) model, we assume that:

$$b_{kj} = \gamma_k + \delta_{kj},$$

where γ_k represnts overall effect of β_k

δ_j represents the difference in the coefficient between sub-population j and the overall coefficient, where $E[\delta_{kj}] = 0$

The Random Coefficient Model

- Suppose

$$y_{ij} \sim N(m_{ij}, t_j)$$

then

$$m_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + ... + \beta_{kj} X_{kj}$$

- For a **random coefficient** (hierarchical) model, we assume that:

$$b_{kj} = \gamma_k + \delta_{kj},$$

where γ_k represnts overall effect of β_k

δ_j represents the difference in the coefficient between sub-population j and the overall coefficient, where $E[\delta_{kj}] = 0$

Random Coefficient Model and its Prior

- Suppose that $y_{ij} \sim N(m_{ij}, t_j)$

$$m_{ij} = b_{0j} + b_{1j}X_{1j} + ... + b_{mj}X_{mj}$$

then

$$m_{ij} = (\gamma_0 + \delta_{0j}) + (\gamma_1 + \delta_{1j})X_{1j} + ... + (\gamma_1 + \delta_{1j})X_{mj}$$

- We shall assume that $t_j \sim Gamma(0.001, 0.001)$ for all j

Random Coefficient Model and its Prior

- Suppose that $y_{ij} \sim N(m_{ij}, t_j)$

$$m_{ij} = b_{0j} + b_{1j}X_{1j} + ... + b_{mj}X_{mj}$$

then

$$m_{ij} = (\gamma_0 + \delta_{0j}) + (\gamma_1 + \delta_{1j})X_{1j} + ... + (\gamma_1 + \delta_{1j})X_{mj}$$

- We shall assume that $t_j \sim Gamma(0.001, 0.001)$ for all j

Random Coefficient Model and its Prior

- Two basic strategies for defining priors for the coefficients:
  1. Specify priors for both γ_k and δ_{kj} as follows:
     $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\delta_{kj} \sim N(0, \tau_k)$ where
     $\tau_k \sim Gamma(0.001, 0.001)$
  2. Use "Hierarchical-centering" as follows: $\beta_{kj} \sim N(\gamma_k, \tau_k)$ where
     $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and
     $\tau_k \sim Gamma(0.001, 0.001)$
- Method 2 improves MCMC markedly in some cases (see Gilks and Roberts, "Strategies for improving MCMC" in MCMC in Practice)

Random Coefficient Model and its Prior

- Two basic strategies for defining priors for the coefficients:
  1. Specify priors for both γ_k and δ_{kj} as follows: $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\delta_{kj} \sim N(0, \tau_k)$ where $\tau_k \sim Gamma(0.001, 0.001)$
  2. Use "Hierarchical-centering" as follows: $\beta_{kj} \sim N(\gamma_k, \tau_k)$ where $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\tau_k \sim Gamma(0.001, 0.001)$
- Method 2 improves MCMC markedly in some cases (see Gilks and Roberts, "Strategies for improving MCMC" in MCMC in Practice)

Random Coefficient Model and its Prior

- Two basic strategies for defining priors for the coefficients:
  1. Specify priors for both γ_k and δ_{kj} as follows:
     $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\delta_{kj} \sim N(0, \tau_k)$ where
     $\tau_k \sim Gamma(0.001, 0.001)$
  2. Use "Hierarchical-centering" as follows: $\beta_{kj} \sim N(\gamma_k, \tau_k)$ where
     $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and
     $\tau_k \sim Gamma(0.001, 0.001)$
- Method 2 improves MCMC markedly in some cases (see Gilks and Roberts, "Strategies for improving MCMC" in MCMC in Practice)

Example from last Module

- **Dependent variable**:

  Percentage of seats won by the Democratic Party in the House of Representatives in state i in election t.

- **Independent variable**:
- Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

- Control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- **Dependent variable**:

  Percentage of seats won by the Democratic Party in the House of Representatives in state i in election t.

- **Independent variable**:
- Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

- Control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- **Dependent variable**:

  Percentage of seats won by the Democratic Party in the House of Representatives in state i in election t.

- **Independent variable**:
- Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

- Control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

Hierarchical Binomial Linear Regression Model using the logit link

- The model takes the following form:

$$y_i \sim Bernoulli(\theta_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the logit link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Binomial Linear Regression Model using the logit link

- The model takes the following form:

$$y_i \sim Bernoulli(\theta_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the logit link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Binomial Linear Regression Model using the logit link

- The model takes the following form:

$$y_i \sim Bernoulli(\theta_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the logit link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Binomial Linear Regression Model using the logit link

- The model takes the following form:

$$y_i \sim Bernoulli(\theta_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the logit link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Binomial Linear Regression Model using the logit link

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Binomial Linear Regression Model using the logit link

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Binomial Linear Regression Model using the logit link

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Binomial Linear Regression Model using the logit link

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Binomial Linear Regression Model using the logit link

- It is difficult to have default parameters for the priors on the precision matrix for the random effects.
- When fitting one of these models, it is of utmost importance to choose a prior that reflects your prior beliefs about the random effects.
- Using the dwish and rwish functions might be useful in choosing these values.

Hierarchical Binomial Linear Regression Model using the logit link

- It is difficult to have default parameters for the priors on the precision matrix for the random effects.

- When fitting one of these models, it is of utmost importance to choose a prior that reflects your prior beliefs about the random effects.

- Using the dwish and rwish functions might be useful in choosing these values.

Hierarchical Binomial Linear Regression Model using the logit link

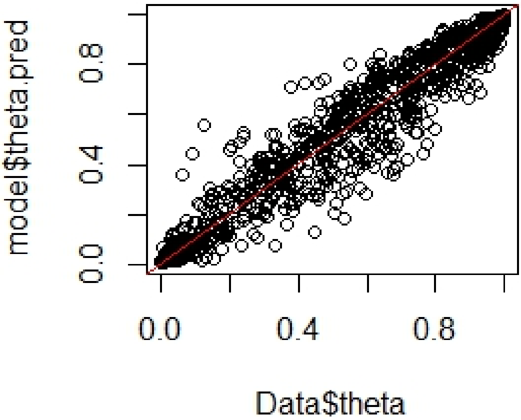
- It is difficult to have default parameters for the priors on the precision matrix for the random effects.
- When fitting one of these models, it is of utmost importance to choose a prior that reflects your prior beliefs about the random effects.
- Using the dwish and rwish functions might be useful in choosing these values.

Implementation in R

- In MCMCpack R-package the MCMChlogit implement the Hierarchical Binomial Linear Regression Model with the logit link function

```
#== Call to MCMChlogit
model <- MCMChlogit(fixed=Y~X1+X2, random=~X1+X2, group="species",
              data=Data, burnin=1000, mcmc=1000, thin=1,verbose=1,
              seed=NA, beta.start=0, sigma2.start=1,
              Vb.start=1, mubeta=0, Vbeta=1.0E6,
              r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)
```

Implementation in R



Hierarchical Poisson Linear Regression Model using the log link function

- The model takes the following form:

$$y_i \sim Poisson(\lambda_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the log link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Poisson Linear Regression Model using the log link function

- The model takes the following form:

$$y_i \sim Poisson(\lambda_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the log link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Poisson Linear Regression Model using the log link function

- The model takes the following form:

$$y_i \sim Poisson(\lambda_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the log link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Poisson Linear Regression Model using the log link function

- The model takes the following form:

$$y_i \sim Poisson(\lambda_i)$$

- With latent variables $l(\theta)$, $l(.)$ being the log link function:

$$l(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

Hierarchical Poisson Linear Regression Model using the log link function

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Poisson Linear Regression Model using the log link function

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

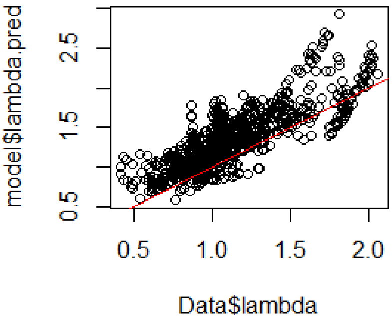
$$V_b \sim IWishart(r, rR)$$

Implementation in R

- In MCMCpack R-package the MCMChpoisson implement the Hierarchical Poisson Linear Regression Model with the log link function

```
#== Call to MCMChpoisson
model <- MCMChpoisson(fixed=Y~X1+X2, random=~X1+X2, group="species",
            data=Data, burnin=500, mcmc=1000, thin=1,verbose=1,
            seed=NA, beta.start=0, sigma2.start=1,
            Vb.start=1, mubeta=0, Vbeta=1.0E6,
            r=3, R=diag(c(0.1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)
```

Implementation in R



λ's are overestimated by the models

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Bayesian Generalized Linear Models

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Generalized Linear Model (GLM)
2. Bayesian setup of GLM
3. R-implementation

1. Generalized Linear Model (GLM)
2. Bayesian setup of GLM
3. R-implementation

Outline

1. Generalized Linear Model (GLM)
2. Bayesian setup of GLM
3. R-implementation

- In general, statistical models contain both systematic and random components.

- For the standard linear model, we assume that Y (the dep. var) is a vector of random variables whose components are independently distributed with mean μ

- μ represents the systematic component of the model (the expected value of Y) and is assumed to be a linear function of explanatory variables X and parameters b.

- The random part of the model (the unexplainable error terms) are assumed to be independent with constant error variance.

Review of GLMS

- In general, statistical models contain both systematic and random components.

- For the standard linear model, we assume that Y (the dep. var) is a vector of random variables whose components are independently distributed with mean μ

- μ represents the systematic component of the model (the expected value of Y) and is assumed to be a linear function of explanatory variables X and parameters b.

- The random part of the model (the unexplainable error terms) are assumed to be independent with constant error variance.

Review of GLMS

- In general, statistical models contain both systematic and random components.

- For the standard linear model, we assume that Y (the dep. var) is a vector of random variables whose components are independently distributed with mean μ

- μ represents the systematic component of the model (the expected value of Y) and is assumed to be a linear function of explanatory variables X and parameters b.

- The random part of the model (the unexplainable error terms) are assumed to be independent with constant error variance.

e-Pathshala पाठशाला MHRD Govt. of India

- In general, statistical models contain both systematic and random components.

- For the standard linear model, we assume that Y (the dep. var) is a vector of random variables whose components are independently distributed with mean μ

- μ represents the systematic component of the model (the expected value of Y) and is assumed to be a linear function of explanatory variables X and parameters b.

- The random part of the model (the unexplainable error terms) are assumed to be independent with constant error variance.

- GLM has three important components:
  1. **random component**: each observation or component of Y has an independent normal distribution with $E(Y) = \mu$ and constant variance σ^2.
  2. **systematic component**: covariates X produce a linear predictor $\eta = X\beta$
  3. **link function**: The link between the random and the systematic components $\mu = \eta$
- For the normal linear model, the link states that the linear predictor $\eta = X\beta$ is identical to the expected value of the random component.
- However, more generally, $\eta_i = g(\mu_i)$, where $g()$ is called the link function.

- GLM has three important components:
  1. **random component**: each observation or component of Y has an independent normal distribution with $E(Y) = \mu$ and constant variance σ^2.
  2. **systematic component**: covariates X produce a linear predictor $\eta = X\beta$
  3. **link function**: The link between the random and the systematic components $\mu = \eta$
- For the normal linear model, the link states that the linear predictor $\eta = X\beta$ is identical to the expected value of the random component.
- However, more generally, $\eta_i = g(\mu_i)$, where $g()$ is called the link function.

- GLM has three important components:
  1. **random component**: each observation or component of Y has an independent normal distribution with $E(Y) = \mu$ and constant variance σ^2.
  2. **systematic component**: covariates X produce a linear predictor $\eta = X\beta$
  3. **link function**: The link between the random and the systematic components $\mu = \eta$
- For the normal linear model, the link states that the linear predictor $\eta = X\beta$ is identical to the expected value of the random component.
- However, more generally, $\eta_i = g(\mu_i)$, where $g()$ is called the link function.

Review of GLM

- GLM has three important components:
  1. **random component**: each observation or component of Y has an independent normal distribution with $E(Y) = \mu$ and constant variance σ^2.
  2. **systematic component**: covariates X produce a linear predictor $\eta = X\beta$
  3. **link function**: The link between the random and the systematic components $\mu = \eta$
- For the normal linear model, the link states that the linear predictor $\eta = X\beta$ is identical to the expected value of the random component.
- However, more generally, $\eta_i = g(\mu_i)$, where $g()$ is called the link function.

- GLM has three important components:
  1. **random component**: each observation or component of Y has an independent normal distribution with $E(Y) = \mu$ and constant variance σ^2.

  2. **systematic component**: covariates X produce a linear predictor $\eta = X\beta$

  3. **link function**: The link between the random and the systematic components $\mu = \eta$
- For the normal linear model, the link states that the linear predictor $\eta = X\beta$ is identical to the expected value of the random component.

- However, more generally, $\eta_i = g(\mu_i)$, where $g()$ is called the link function.

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- General linear models extend the above setup to the case where:
  1. the random component follows a distribution other than the normal
  2. the link function is a function other than that given above.
- Common distributions other than the normal for the random component include...
  1. Poisson
  2. Bernoulli / Binomial
  3. Weibull (for modeling duration)
  4. Multinomial

- The link function relates the linear predictor η to an observation y.

- In classical linear models, where the dependent variable is normal, we use the *identity link*. Since the expected value of the linear predictor can take any value for the normal distribution, the identity link makes sense.

- For a Poisson (count) model, $\mu > 0$, so the identity link is less attractive because the linear predictor η can be negative while μ cannot.

- So, for Poisson models we typically use the log link, $\eta = \log(\mu)$, or its inverse $\mu = \exp\{\eta\}$.

- The link function relates the linear predictor η to an observation y.

- In classical linear models, where the dependent variable is normal, we use the *identity link*. Since the expected value of the linear predictor can take any value for the normal distribution, the identity link makes sense.

- For a Poisson (count) model, $\mu > 0$, so the identity link is less attractive because the linear predictor η can be negative while μ cannot.

- So, for Poisson models we typically use the log link, $\eta = \log(\mu)$, or its inverse $\mu = \exp\{\eta\}$.

- The link function relates the linear predictor η to an observation y.
- In classical linear models, where the dependent variable is normal, we use the *identity link*. Since the expected value of the linear predictor can take any value for the normal distribution, the identity link makes sense.
- For a Poisson (count) model, $\mu > 0$, so the identity link is less attractive because the linear predictor η can be negative while μ cannot.
- So, for Poisson models we typically use the log link, $\eta = \log(\mu)$, or its inverse $\mu = \exp\{\eta\}$.

- The link function relates the linear predictor η to an observation y.
- In classical linear models, where the dependent variable is normal, we use the *identity link*. Since the expected value of the linear predictor can take any value for the normal distribution, the identity link makes sense.
- For a Poisson (count) model, $\mu > 0$, so the identity link is less attractive because the linear predictor η can be negative while μ cannot.
- So, for Poisson models we typically use the log link, $\eta = \log(\mu)$, or its inverse $\mu = \exp\{\eta\}$.

- Link function maps the linear predictor which can take any value along the real line to a set of plausible expected values.
- For a Bernoulli model, $0 < \mu < 1$, so the identity link is unattractive. So, for Bernoulli models, we typically use the logit link, $\eta = \log(\mu/(1-\mu))$

- Link function maps the linear predictor which can take any value along the real line to a set of plausible expected values.
- For a Bernoulli model, $0 < \mu < 1$, so the identity link is unattractive. So, for Bernoulli models, we typically use the logit link, $\eta = \log(\mu/(1-\mu))$

Bayesian Setup of the GLM

- The Bayesian setup for the GLM is a very natural extension of the framework we have used for regression models.
  1. Specify the probability distribution for the dependent variable in your model.
  $$y_i \sim N(\mu_i, t)$$
  2. Define the linear predictor for your model.
  $$b_1 + b_2 X_i$$
  3. Choose the link function that maps from the linear predictor η to a set of plausible expected values for μ
  $$\mu_i = b_1 + b_2 X_i$$
  4. Choose priors for all of the parameters in your model.
  $$b_i \sim N(0, 0.001) \text{ and } t \sim Gamma(0.1, 0.1)$$

Bayesian Setup of the GLM

- The Bayesian setup for the GLM is a very natural extension of the framework we have used for regression models.
  1. Specify the probability distribution for the dependent variable in your model.

$$y_i \sim N(\mu_i, t)$$

  2. Define the linear predictor for your model.

$$b_1 + b_2 X_i$$

  3. Choose the link function that maps from the linear predictor η to a set of plausible expected values for μ

$$\mu_i = b_1 + b_2 X_i$$

  4. Choose priors for all of the parameters in your model.

$$b_i \sim N(0, 0.001) \text{ and } t \sim Gamma(0.1, 0.1)$$

Bayesian Setup of the GLM

- The Bayesian setup for the GLM is a very natural extension of the framework we have used for regression models.
  1. Specify the probability distribution for the dependent variable in your model.

     $$y_i \sim N(\mu_i, t)$$

  2. Define the linear predictor for your model.

     $$b_1 + b_2 X_i$$

  3. Choose the link function that maps from the linear predictor η to a set of plausible expected values for μ

     $$\mu_i = b_1 + b_2 X_i$$

  4. Choose priors for all of the parameters in your model.

     $b_i \sim N(0, 0.001)$ and $t \sim Gamma(0.1, 0.1)$

Bayesian Setup of the GLM

- The Bayesian setup for the GLM is a very natural extension of the framework we have used for regression models.
  1. Specify the probability distribution for the dependent variable in your model.
  $$y_i \sim N(\mu_i, t)$$
  2. Define the linear predictor for your model.
  $$b_1 + b_2 X_i$$
  3. Choose the link function that maps from the linear predictor η to a set of plausible expected values for μ
  $$\mu_i = b_1 + b_2 X_i$$
  4. Choose priors for all of the parameters in your model.
  $$b_i \sim N(0, 0.001) \text{ and } t \sim Gamma(0.1, 0.1)$$

Bayesian Setup of the GLM

- The Bayesian setup for the GLM is a very natural extension of the framework we have used for regression models.
  1. Specify the probability distribution for the dependent variable in your model.
  $$y_i \sim N(\mu_i, t)$$
  2. Define the linear predictor for your model.
  $$b_1 + b_2 X_i$$
  3. Choose the link function that maps from the linear predictor η to a set of plausible expected values for μ
  $$\mu_i = b_1 + b_2 X_i$$
  4. Choose priors for all of the parameters in your model.
  $$b_i \sim N(0, 0.001) \text{ and } t \sim Gamma(0.1, 0.1)$$

- Suppose that $y_i \sim Bernoulli(p_i)$,

- To ensure that $0 < p_i < 1$, we use the logit transformation so,

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

  we assume $\beta_j \sim N(0, 0.001)$ for all j.

- notice-there is no prior distribution for the variance of y because there is not a parameter in the Bernoulli distribution for variance.

- Suppose that $y_i \sim Bernoulli(p_i)$,
- To ensure that $0 < p_i < 1$, we use the logit transformation so,

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

  we assume $\beta_j \sim N(0, 0.001)$ for all j.

- notice-there is no prior distribution for the variance of y because there is not a parameter in the Bernoulli distribution for variance.

- Suppose that $y_i \sim Bernoulli(p_i)$,

- To ensure that $0 < p_i < 1$, we use the logit transformation so,

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

we assume $\beta_j \sim N(0, 0.001)$ for all j.

- notice-there is no prior distribution for the variance of y because there is not a parameter in the Bernoulli distribution for variance.

- Thus, the joint posterior distribution of the parameters is given by the following expression:

$$
\begin{aligned}
p(b_0, b_1, b_2, ..., b_k|y, X) &\propto p(b_0, b_1, b_2, ..., b_k) \\
&\quad \prod_{i=1}^{n} p(y_i|b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\propto p(b_0, b_1, b_2, ..., b_k) \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \\
&\propto p(b_0, b_1, b_2, ..., b_k) \\
&\quad \prod_{i=1}^{n} logit^{-1}(b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\quad logit^{-1}(1 - (b_0 + b_1 X_{1i} + ... + b_k X_{ki}))
\end{aligned}
$$

- Thus, the joint posterior distribution of the parameters is given by the following expression:

$$
\begin{aligned}
p(b_0, b_1, b_2, ..., b_k | y, X) &\propto p(b_0, b_1, b_2, ..., b_k) \\
&\quad \prod_{i=1}^{n} p(y_i | b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\propto p(b_0, b_1, b_2, ..., b_k) \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \\
&\propto p(b_0, b_1, b_2, ..., b_k) \\
&\quad \prod_{i=1}^{n} logit^{-1}(b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\quad logit^{-1}(1 - (b_0 + b_1 X_{1i} + ... + b_k X_{ki}))
\end{aligned}
$$

- Thus, the joint posterior distribution of the parameters is given by the following expression:

$$
\begin{aligned}
p(b_0, b_1, b_2, ..., b_k | y, X) \quad &\propto \quad p(b_0, b_1, b_2, ..., b_k) \\
&\quad\quad \prod_{i=1}^{n} p(y_i | b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\propto \quad p(b_0, b_1, b_2, ..., b_k) \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \\
&\propto \quad p(b_0, b_1, b_2, ..., b_k) \\
&\quad\quad \prod_{i=1}^{n} logit^{-1}(b_0 + b_1 X_{1i} + ... + b_k X_{ki}) \\
&\quad\quad logit^{-1}(1 - (b_0 + b_1 X_{1i} + ... + b_k X_{ki}))
\end{aligned}
$$

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

- Choce of prior on regression coefficients:
  1. default improper uniform prior
  2. independent normal prior
  3. multivariate normal prior
  4. Zellner's g-prior
  5. independent Cauchy prior
  6. Lasso or double exponential prior

Application: birtwt dataset

```
> library(MCMCpack)
> ## default improper uniform prior
>    data(birthwt)
>    posterior <- MCMClogit(low~age+as.factor(race)+smoke,
>    round(summary(posterior)$quantiles,digit=3)

                     2.5%    25%    50%    75% 97.5%
(Intercept)        -2.724 -1.601 -1.007 -0.319 0.872
age                -0.110 -0.061 -0.037 -0.015 0.026
as.factor(race)2    0.070  0.698  1.039  1.383 2.017
as.factor(race)3    0.305  0.793  1.061  1.364 1.950
smoke               0.386  0.854  1.141  1.395 1.888

>
```

Application: birtwt dataset

```
> library(MCMCpack)
> ##  multivariate normal prior
>    data(birthwt)
>    posterior <- MCMClogit(low~age+as.factor(race)+smoke,
+                                  , data=birthwt)
>    round(summary(posterior)$quantiles,digit=3)

                     2.5%    25%    50%    75% 97.5%
(Intercept)        -2.799 -1.691 -1.064 -0.409 0.790
age                -0.107 -0.059 -0.035 -0.013 0.030
as.factor(race)2    0.036  0.699  1.023  1.372 2.017
as.factor(race)3    0.291  0.802  1.085  1.369 1.920
smoke               0.380  0.874  1.132  1.400 1.924

>
```

Poisson Regression

- Suppose that $y_i \sim Poisson(\lambda_i)$
- To ensure that $\lambda_i > 0$, we use the log link function,

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- $b_j \sim N(0, 0.001)$ for all j

Poisson Regression

- Suppose that $y_i \sim Poisson(\lambda_i)$
- To ensure that $\lambda_i > 0$, we use the log link function,

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- $b_j \sim N(0, 0.001)$ for all j

- Suppose that $y_i \sim Poisson(\lambda_i)$

- To ensure that $\lambda_i > 0$, we use the log link function,

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- $b_j \sim N(0, 0.001)$ for all j

MCMC Implementation for Poisson Regression

```
>   counts <- c(18,17,15,20,10,20,25,13,12)
>    outcome <- gl(3,1,9)
>    treatment <- gl(3,3)
>    posterior <- MCMCpoisson(counts ~ outcome + treatment)
>    round(summary(posterior)$quantiles,digit=3)
```

```
              2.5%    25%    50%    75%  97.5%
(Intercept)  2.662  2.906  3.030  3.154  3.355
outcome2    -0.828 -0.578 -0.457 -0.319 -0.059
outcome3    -0.676 -0.412 -0.286 -0.157  0.087
treatment2  -0.388 -0.138  0.000  0.143  0.394
treatment3  -0.400 -0.146 -0.003  0.130  0.388
```

- The model takes the following form:

$$y_i \sim Multinomial(\pi_i)$$

where:

$$\pi_{ij} = \exp(x'_{ij}\beta)/[\sum_{k=1}^{p} \exp(x'_{ik}\beta)]$$

We assume a multivariate Normal prior on beta:

$$\beta \; N(b_0, B_0^{-1})$$

MCMC for Multinomial Logit Regression

```
> data(Nethvote)
>  post<- MCMCmnl(vote ~
+                 relig + class + income + educ + age + ur
+                 baseline="D66", mcmc.method="IndMH", B0=
+                 verbose=0, mcmc=1000, thin=1, tune=0.5,
+                 data=Nethvote)
Calculating MLEs and large sample var-cov matrix.
This may take a moment...
Inverting Hessian to get large sample var-cov matrix.
>   head(round(summary(post)$quantile,digit=3))
                   2.5%    25%    50%    75% 97.5%
(Intercept).CDA  -0.773 -0.320 -0.075  0.171 0.626
(Intercept).PvdA  1.970  2.312  2.475  2.677 3.028
(Intercept).VVD  -1.181 -0.744 -0.507 -0.220 0.129
relig.CDA         1.947  2.331  2.430  2.540 2.916
relig.PvdA       -0.118  0.089  0.220  0.364 0.659
```

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

Subject: Statistics

Paper: Statistical Inference
Module: Missing Data Models

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Introduction to missing data models

2. Types of missing data

3. Corrections for missing data in theory

4. Missing data in practice

Outline

1. Introduction to missing data models
2. Types of missing data
3. Corrections for missing data in theory
4. Missing data in practice

Outline

1. Introduction to missing data models
2. Types of missing data
3. Corrections for missing data in theory
4. Missing data in practice

Outline

1. Introduction to missing data models
2. Types of missing data
3. Corrections for missing data in theory
4. Missing data in practice

- It is well known that on average about half of the respondents to surveys in political science research do not answer on or more of the questions used in the analysis.
- Almost all analysts "contaminate" their data by filling in educated guesses for some of these questions (e.g. for party identification questions, don't know = independent).
- Even if these guesses are correct on average, filling in missing cells of our data matrix in this way biases our regression coefficients' standard errors downwards.

- It is well known that on average about half of the respondents to surveys in political science research do not answer on or more of the questions used in the analysis.
- Almost all analysts "contaminate" their data by filling in educated guesses for some of these questions (e.g. for party identification questions, don't know = independent).
- Even if these guesses are correct on average, filling in missing cells of our data matrix in this way biases our regression coefficients' standard errors downwards.

- It is well known that on average about half of the respondents to surveys in political science research do not answer on or more of the questions used in the analysis.
- Almost all analysts "contaminate" their data by filling in educated guesses for some of these questions (e.g. for party identification questions, don't know = independent).
- Even if these guesses are correct on average, filling in missing cells of our data matrix in this way biases our regression coefficients' standard errors downwards.

- When educated guesses are not possible, the standard "remedy" is listwise deletion of missing data, eliminating entire observations in a wholesale manner.
- Valuable information is lost, and severe selection bias effects are possible.

- When educated guesses are not possible, the standard "remedy" is listwise deletion of missing data, eliminating entire observations in a wholesale manner.
- Valuable information is lost, and severe selection bias effects are possible.

Some notation

- Let D denote the data matrix, where D includes independent and dependent variables. $D = \{X, y\}$.
- We assume that some elements of the data matrix are missing.
- Let M denote the missingness indicator matrix with the same dimensions of D. Each element of M is a one or zero that indicates whether or not an element of D is missing.
- $D_{ij} = 0$ indicates the i^{th} observation for j^{th} variable is missing but that the data could be observed
- $D_{ij} = 1$ means that piece of data is present.

Some notation

- Let D denote the data matrix, where D includes independent and dependent variables. $D = \{X, y\}$.
- We assume that some elements of the data matrix are missing.
- Let M denote the missingness indicator matrix with the same dimensions of D. Each element of M is a one or zero that indicates whether or not an element of D is missing.
- $D_{ij} = 0$ indicates the i^{th} observation for j^{th} variable is missing but that the data could be observed
- $D_{ij} = 1$ means that piece of data is present.

Some notation

- Let D denote the data matrix, where D includes independent and dependent variables. $D = \{X, y\}$.
- We assume that some elements of the data matrix are missing.
- Let M denote the missingness indicator matrix with the same dimensions of D. Each element of M is a one or zero that indicates whether or not an element of D is missing.
- $D_{ij} = 0$ indicates the i^{th} observation for j^{th} variable is missing but that the data could be observed
- $D_{ij} = 1$ means that piece of data is present.

- Let D denote the data matrix, where D includes independent and dependent variables. $D = \{X, y\}$.
- We assume that some elements of the data matrix are missing.
- Let M denote the missingness indicator matrix with the same dimensions of D. Each element of M is a one or zero that indicates whether or not an element of D is missing.
- $D_{ij} = 0$ indicates the i^{th} observation for j^{th} variable is missing but that the data could be observed
- $D_{ij} = 1$ means that piece of data is present.

Some notation

- Let D denote the data matrix, where D includes independent and dependent variables. $D = \{X, y\}$.

- We assume that some elements of the data matrix are missing.

- Let M denote the missingness indicator matrix with the same dimensions of D. Each element of M is a one or zero that indicates whether or not an element of D is missing.

- $D_{ij} = 0$ indicates the i^{th} observation for j^{th} variable is missing but that the data could be observed

- $D_{ij} = 1$ means that piece of data is present.

- **Comment**: it is possible that data cannot be observed. Sometimes a "don't know" really means that the respondent has no basis on which to provide an answer.

- Finally, let D_{obs} and D_{mis} denote the observed and missing parts of the D. $D = \{D_{obs}, D_{mis}\}$.

Comments

- **Comment**: it is possible that data cannot be observed. Sometimes a "don't know" really means that the respondent has no basis on which to provide an answer.

- Finally, let D_{obs} and D_{mis} denote the observed and missing parts of the D. $D = \{D_{obs}, D_{mis}\}$.

Three Types of Missingness

- **Missing Completely at Random**: if the data are missing completely at random then missing values cannot be predicted any better with the information in D, observed or not.

- Formally, M is independent of D. So, $P(M|D) = P(M)$.

- A process is missing completely at random if, say, an individual's decides whether to answer survey questions on the basis of coin flips.

- If independent are more likely to decline to answer a vote preference or party id question, then the data are not missing completely at random.
- In the unlikely event that the process is missing completely at random, then inferences based on listwise deletion are unbiased, but inefficient because we have lost some cases.

- **Missing Completely at Random**: if the data are missing completely at random then missing values cannot be predicted any better with the information in D, observed or not.

- Formally, M is independent of D. So, $P(M|D) = P(M)$.

- A process is missing completely at random if, say, an individual's decides whether to answer survey questions on the basis of coin flips.

- If independent are more likely to decline to answer a vote preference or party id question, then the data are not missing completely at random.
- In the unlikely event that the process is missing completely at random, then inferences based on listwise deletion are unbiased, but inefficient because we have lost some cases.

- **Missing Completely at Random**: if the data are missing completely at random then missing values cannot be predicted any better with the information in D, observed or not.

- Formally, M is independent of D. So, $P(M|D) = P(M)$.

- A process is missing completely at random if, say, an individual's decides whether to answer survey questions on the basis of coin flips.

- If independent are more likely to decline to answer a vote preference or party id question, then the data are not missing completely at random.
- In the unlikely event that the process is missing completely at random, then inferences based on listwise deletion are unbiased, but inefficient because we have lost some cases.

- **Missing Completely at Random**: if the data are missing completely at random then missing values cannot be predicted any better with the information in D, observed or not.

- Formally, M is independent of D. So, $P(M|D) = P(M)$.

- A process is missing completely at random if, say, an individual's decides whether to answer survey questions on the basis of coin flips.

- If independent are more likely to decline to answer a vote preference or party id question, then the data are not missing completely at random.
- In the unlikely event that the process is missing completely at random, then inferences based on listwise deletion are unbiased, but inefficient because we have lost some cases.

- **Missing Completely at Random**: if the data are missing completely at random then missing values cannot be predicted any better with the information in D, observed or not.

- Formally, M is independent of D. So, $P(M|D) = P(M)$.

- A process is missing completely at random if, say, an individual's decides whether to answer survey questions on the basis of coin flips.

- If independent are more likely to decline to answer a vote preference or party id question, then the data are not missing completely at random.
- In the unlikely event that the process is missing completely at random, then inferences based on listwise deletion are unbiased, but inefficient because we have lost some cases.

Three Types of Missingness

- **Missing at Random**: if the data are missing at random then the probability that a cell is missing may depend on D_{obs}, but after controlling for D_{obs} that probability must be independent of D_{mis}.
- In other words, the process that determines whether or not a cell is missing should not depend on the values in the cell.
- Formally, M is independent of D_{mis} : $P(M|D) = P(M|D_{obs})$
- For example, if Democratic identifiers are more likely to refuse to answer the vote choice question, then the process is missing at random so long as party id is a question to which at least some people respond.
- If data is missing at random, then inferences based on listwise deletion will be biased and inefficient.

Three Types of Missingness

- **Missing at Random**: if the data are missing at random then the probability that a cell is missing may depend on D_{obs}, but after controlling for D_{obs} that probability must be independent of D_{mis}.
- In other words, the process that determines whether or not a cell is missing should not depend on the values in the cell.
- Formally, M is independent of D_{mis} : $P(M|D) = P(M|D_{obs})$
- For example, if Democratic identifiers are more likely to refuse to answer the vote choice question, then the process is missing at random so long as party id is a question to which at least some people respond.
- If data is missing at random, then inferences based on listwise deletion will be biased and inefficient.

- **Missing at Random**: if the data are missing at random then the probability that a cell is missing may depend on D_{obs}, but after controlling for D_{obs} that probability must be independent of D_{mis}.
- In other words, the process that determines whether or not a cell is missing should not depend on the values in the cell.
- Formally, M is independent of D_{mis} : $P(M|D) = P(M|D_{obs})$
- For example, if Democratic identifiers are more likely to refuse to answer the vote choice question, then the process is missing at random so long as party id is a question to which at least some people respond.
- If data is missing at random, then inferences based on listwise deletion will be biased and inefficient.

- **Missing at Random**: if the data are missing at random then the probability that a cell is missing may depend on D_{obs}, but after controlling for D_{obs} that probability must be independent of D_{mis}.
- In other words, the process that determines whether or not a cell is missing should not depend on the values in the cell.
- Formally, M is independent of D_{mis} : $P(M|D) = P(M|D_{obs})$
- For example, if Democratic identifiers are more likely to refuse to answer the vote choice question, then the process is missing at random so long as party id is a question to which at least some people respond.
- If data is missing at random, then inferences based on listwise deletion will be biased and inefficient.

- **Missing at Random**: if the data are missing at random then the probability that a cell is missing may depend on D_{obs}, but after controlling for D_{obs} that probability must be independent of D_{mis}.
- In other words, the process that determines whether or not a cell is missing should not depend on the values in the cell.
- Formally, M is independent of D_{mis} : $P(M|D) = P(M|D_{obs})$
- For example, if Democratic identifiers are more likely to refuse to answer the vote choice question, then the process is missing at random so long as party id is a question to which at least some people respond.
- If data is missing at random, then inferences based on listwise deletion will be biased and inefficient.

- **Non-Ignorable**:if the probability that a cell is missing depends on the unobserved value of the missing response, then the process is non-ignorable.
- Formally, $P(M|D)$ cannot be simplified.
- A standard example is individuals' responses to income questions, where high income people are more likely to refuse to answer survey questions about income and other variables in the data set cannot predict which respondents have high income.
- If your missing data is non-ignorable, then inferences based on listwise deletion will be biased and inefficient (and multiple imputation algorithms that we will talk about wont be of much aid).

- **Non-Ignorable**:if the probability that a cell is missing depends on the unobserved value of the missing response, then the process is non-ignorable.

- Formally, $P(M|D)$ cannot be simplified.

- A standard example is individuals' responses to income questions, where high income people are more likely to refuse to answer survey questions about income and other variables in the data set cannot predict which respondents have high income.

- If your missing data is non-ignorable, then inferences based on listwise deletion will be biased and inefficient (and multiple imputation algorithms that we will talk about wont be of much aid).

- **Non-Ignorable**:if the probability that a cell is missing depends on the unobserved value of the missing response, then the process is non-ignorable.
- Formally, $P(M|D)$ cannot be simplified.
- A standard example is individuals' responses to income questions, where high income people are more likely to refuse to answer survey questions about income and other variables in the data set cannot predict which respondents have high income.
- If your missing data is non-ignorable, then inferences based on listwise deletion will be biased and inefficient (and multiple imputation algorithms that we will talk about wont be of much aid).

Three Types of Missingness

- **Non-Ignorable**:if the probability that a cell is missing depends on the unobserved value of the missing response, then the process is non-ignorable.

- Formally, $P(M|D)$ cannot be simplified.

- A standard example is individuals' responses to income questions, where high income people are more likely to refuse to answer survey questions about income and other variables in the data set cannot predict which respondents have high income.

- If your missing data is non-ignorable, then inferences based on listwise deletion will be biased and inefficient (and multiple imputation algorithms that we will talk about wont be of much aid).

Multiple Imputation of Missing Data

- Multiple imputation involves imputing m values for each missing item and creating m completed data sets.

- Across each of the m data sets:
  if $M_{ij} = 1$ then $D_{ij} =$ observed data
  if $M_{ij} = 0$ then $D_{ij} =$ an imputed value

- The imputed values for the data set are based on our guesses about the value of M_{ij} together with summaries of our uncertainty regarding the missing values.

- For each imputed data set, you perform whatever statistical analysis you normally would. Then, you average your results over the m computed analyses.

Multiple Imputation of Missing Data

- Multiple imputation involves imputing m values for each missing item and creating m completed data sets.

- Across each of the m data sets:
  if $M_{ij} = 1$ then $D_{ij} =$ observed data
  if $M_{ij} = 0$ then $D_{ij} =$ an imputed value

- The imputed values for the data set are based on our guesses about the value of M_{ij} together with summaries of our uncertainty regarding the missing values.

- For each imputed data set, you perform whatever statistical analysis you normally would. Then, you average your results over the m computed analyses.

Multiple Imputation of Missing Data

- Multiple imputation involves imputing m values for each missing item and creating m completed data sets.

- Across each of the m data sets:

  if $M_{ij} = 1$ then $D_{ij} =$ observed data

  if $M_{ij} = 0$ then $D_{ij} =$ an imputed value

- The imputed values for the data set are based on our guesses about the value of M_{ij} together with summaries of our uncertainty regarding the missing values.

- For each imputed data set, you perform whatever statistical analysis you normally would. Then, you average your results over the m computed analyses.

Multiple Imputation of Missing Data

- Multiple imputation involves imputing m values for each missing item and creating m completed data sets.

- Across each of the m data sets:

  if $M_{ij} = 1$ then $D_{ij} =$ observed data

  if $M_{ij} = 0$ then $D_{ij} =$ an imputed value

- The imputed values for the data set are based on our guesses about the value of M_{ij} together with summaries of our uncertainty regarding the missing values.

- For each imputed data set, you perform whatever statistical analysis you normally would. Then, you average your results over the m computed analyses.

Calculating Quantities of Interest

- To estimate some quantity of interest Q such as a variable mean or regression coefficient, the overall point estimate q^* of Q is the average of the m separate estimates q_j. That is

$$q^* = \sum_{j=1}^{m} q_j/m$$

- Let $SE(q_j)$ denote the standard error of q_j for data set j and let $S_q^2 = \frac{1}{m-1}\sum_{j=1}^{m}(q_j - q^*)^2$ denote the sample variance across the m point estimates.

- Then the variance of the multiple imputation point estimates is the weighted average of the estimated variances from *within* each completed data set plus the sample variance in the point estimates across the data sets.

Calculating Quantities of Interest

- To estimate some quantity of interest Q such as a variable mean or regression coefficient, the overall point estimate q^* of Q is the average of the m separate estimates q_j. That is

$$q^* = \sum_{j=1}^{m} q_j/m$$

- Let $SE(q_j)$ denote the standard error of q_j for data set j and let $S_q^2 = \frac{1}{m-1}\sum_{j=1}^{m}(q_j - q^*)^2$ denote the sample variance across the m point estimates.

- Then the variance of the multiple imputation point estimates is the weighted average of the estimated variances from *within* each completed data set plus the sample variance in the point estimates across the data sets.

Calculating Quantities of Interest

- To estimate some quantity of interest Q such as a variable mean or regression coefficient, the overall point estimate q^* of Q is the average of the m separate estimates q_j. That is

$$q^* = \sum_{j=1}^{m} q_j/m$$

- Let $SE(q_j)$ denote the standard error of q_j for data set j and let $S_q^2 = \frac{1}{m-1}\sum_{j=1}^{m}(q_j - q^*)^2$ denote the sample variance across the m point estimates.

- Then the variance of the multiple imputation point estimates is the weighted average of the estimated variances from *within* each completed data set plus the sample variance in the point estimates across the data sets.

- The weight is a function of the number of data sets, so that if $m = \infty$, then it would be a straightforward average of the two sources of uncertainty.

- That is

$$SE(q)^2 = (1/m)\sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + 1/m)$$

- The weight is a function of the number of data sets, so that if $m = \infty$, then it would be a straightforward average of the two sources of uncertainty.
- That is

$$SE(q)^2 = (1/m)\sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + 1/m)$$

King, et al's, imputation model

- To implement multiple imputation, we need a statistical model that we can use to sample missing values.

- To use King's AMELIA program (and our most extensive R alternative), we assume that the data are missing at random conditional on the imputation model (where the model is defined to include the variables that provide information about the missing data process.)

- King's missing data program is based on the assumption that all of the variables in our model are jointly multivariate normal density.

  King argues that in most cases, the multivariate normal density is a very good approximation, even if some of our variables are ordinal.

King, et al's, imputation model

- To implement multiple imputation, we need a statistical model that we can use to sample missing values.

- To use King's AMELIA program (and our most extensive R alternative), we assume that the data are missing at random conditional on the imputation model (where the model is defined to include the variables that provide information about the missing data process.)

- King's missing data program is based on the assumption that all of the variables in our model are jointly multivariate normal density.

  King argues that in most cases, the multivariate normal density is a very good approximation, even if some of our variables are ordinal.

King, et al's, imputation model

- To implement multiple imputation, we need a statistical model that we can use to sample missing values.

- To use King's AMELIA program (and our most extensive R alternative), we assume that the data are missing at random conditional on the imputation model (where the model is defined to include the variables that provide information about the missing data process.)

- King's missing data program is based on the assumption that all of the variables in our model are jointly multivariate normal density.

  King argues that in most cases, the multivariate normal density is a very good approximation, even if some of our variables are ordinal.

- Stated more formally, King et al assume that for each observation i $(i = 1, ..., n)$, D_i denotes the vector of values for the p variables including the dependent and independent variables.

- The likelihood function for the complete data set is given by:

$$L(\mu, \Sigma) \propto \prod_{i} N(D_i|\mu, \Sigma)$$

where μ is the vector of p means and Σ is a $p \times p$ dimensional variance-covariance matrix that provides information about how values of the independent variables depend on one another.

King, et al's, imputation model

- Stated more formally, King et al assume that for each observation i $(i = 1, ..., n)$, D_i denotes the vector of values for the p variables including the dependent and independent variables.

- The likelihood function for the complete data set is given by:

$$L(\mu, \Sigma) \propto \prod_i N(D_i | \mu, \Sigma)$$

where μ is the vector of p means and Σ is a $p \times p$ dimensional variance-covariance matrix that provides information about how values of the independent variables depend on one another.

- If we assume that the data are missing at random, then the likelihood function for the observed data are given by:

$$L(\mu, \Sigma | D_{obs}) \propto \prod_i N(D_{i,obs} | \mu_{i,obs}, \Sigma_{i,obs})$$

- King et al note that this will be a tough likelihood to work with because each observation is likely to have a different combination of missing values.

King, et al's, imputation model

- If we assume that the data are missing at random, then the likelihood function for the observed data are given by:

$$L(\mu, \Sigma | D_{obs}) \propto \prod_i N(D_{i,obs} | \mu_{i,obs}, \Sigma_{i,obs})$$

- King et al note that this will be a tough likelihood to work with because each observation is likely to have a different combination of missing values.

King, et al's, imputation model

- The multivariate normal model implies that each missing value is imputed linearly. That is, we can simulate a missing value in the way that we would usually simulate from a regression.

- For example, let D^*_{ij} denote a simulated value of observation i and variable j and let $D_{i,-j}$ denote the vector of values of all [observed] variables in row i except j.

- Then the posterior distribution of the coefficient B_{-j} from a regression of D_j on D_{-j} (which can be calculated from μ and Σ) can be used such that: $D^*_{ij} = D_{i,-j}B_{-j} + e^*_i$.

- An alternative way to think about this is to imagine a multivariate normal distribution. An imputed draw from the multivariate normal distribution is a draw from the slice of the distribution for D_{mis} corresponding to the value of D_{obs}.

King, et al's, imputation model

- The multivariate normal model implies that each missing value is imputed linearly. That is, we can simulate a missing value in the way that we would usually simulate from a regression.

- For example, let D^*_{ij} denote a simulated value of observation i and variable j and let $D_{i,-j}$ denote the vector of values of all [observed] variables in row i except j.

- Then the posterior distribution of the coefficient B_{-j} from a regression of D_j on D_{-j} (which can be calculated from μ and Σ) can be used such that: $D^*_{ij} = D_{i,-j}B_{-j} + e^*_i$.

- An alternative way to think about this is to imagine a multivariate normal distribution. An imputed draw from the multivariate normal distribution is a draw from the slice of the distribution for D_{mis} corresponding to the value of D_{obs}.

- The multivariate normal model implies that each missing value is imputed linearly. That is, we can simulate a missing value in the way that we would usually simulate from a regression.

- For example, let D^*_{ij} denote a simulated value of observation i and variable j and let $D_{i,-j}$ denote the vector of values of all [observed] variables in row i except j.

- Then the posterior distribution of the coefficient B_{-j} from a regression of D_j on D_{-j} (which can be calculated from μ and Σ) can be used such that: $D^*_{ij} = D_{i,-j}B_{-j} + e^*_i$.

- An alternative way to think about this is to imagine a multivariate normal distribution. An imputed draw from the multivariate normal distribution is a draw from the slice of the distribution for D_{mis} corresponding to the value of D_{obs}.

- The multivariate normal model implies that each missing value is imputed linearly. That is, we can simulate a missing value in the way that we would usually simulate from a regression.

- For example, let D^*_{ij} denote a simulated value of observation i and variable j and let $D_{i,-j}$ denote the vector of values of all [observed] variables in row i except j.

- Then the posterior distribution of the coefficient B_{-j} from a regression of D_j on D_{-j} (which can be calculated from μ and Σ) can be used such that: $D^*_{ij} = D_{i,-j}B_{-j} + e^*_i$.

- An alternative way to think about this is to imagine a multivariate normal distribution. An imputed draw from the multivariate normal distribution is a draw from the slice of the distribution for D_{mis} corresponding to the value of D_{obs}.

- Gary King at Harvard has a very easy to use program called AMELIA (now available as R package) that creates data sets with imputed missing values.

- The program is designed to read in a raw data set with missing values and outputs m new data sets with imputed missing values.

- You then run your analyses on the m imputed data sets, take averages of coefficients and standard errors as discussed above, and that is it.

- Gary King at Harvard has a very easy to use program called AMELIA (now available as R package) that creates data sets with imputed missing values.

- The program is designed to read in a raw data set with missing values and outputs m new data sets with imputed missing values.

- You then run your analyses on the m imputed data sets, take averages of coefficients and standard errors as discussed above, and that is it.

- Gary King at Harvard has a very easy to use program called AMELIA (now available as R package) that creates data sets with imputed missing values.
- The program is designed to read in a raw data set with missing values and outputs m new data sets with imputed missing values.
- You then run your analyses on the m imputed data sets, take averages of coefficients and standard errors as discussed above, and that is it.

- Rather than using a small number of multiply imputed data sets, one might incorporate a multiple imputation algorithm into your Gibbs Sampler, taken an independent draw from the multiple imputation data set with each iteration of your program.

- The advantage of this is that the posterior standard errors for regression coefficients already summarize all of the uncertainty about the process for the missing data and the uncertainty about the coefficients themselves.

- The disadvantage is that the imputations will be *much slower and you will have to check for convergence.*

- Rather than using a small number of multiply imputed data sets, one might incorporate a multiple imputation algorithm into your Gibbs Sampler, taken an independent draw from the multiple imputation data set with each iteration of your program.

- The advantage of this is that the posterior standard errors for regression coefficients already summarize all of the uncertainty about the process for the missing data and the uncertainty about the coefficients themselves.

- The disadvantage is that the imputations will be *much slower and you will have to check for convergence.*

- Rather than using a small number of multiply imputed data sets, one might incorporate a multiple imputation algorithm into your Gibbs Sampler, taken an independent draw from the multiple imputation data set with each iteration of your program.

- The advantage of this is that the posterior standard errors for regression coefficients already summarize all of the uncertainty about the process for the missing data and the uncertainty about the coefficients themselves.

- The disadvantage is that the imputations will be *much slower and you will have to check for convergence*.

Missing Data Algorithms

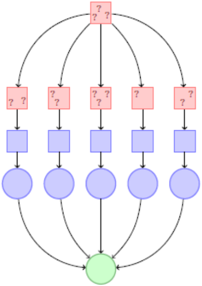


Figure 1: A schematic of our approach to multiple imputation with the EMB algorithm.

- Figure except from Honaker, King and Blackwell (2014)

Amelia : The R package for Missing Data Imputation

- "AMELIA II: A Program for Missing Data", developed by Honaker, King and Blackwell is available as R-package.

- Missing data imputation can easily be handled in Bayesian methods.

- Amelia provides a Bayesian setup to handle missing data imputation

Amelia : The R package for Missing Data Imputation

- "AMELIA II: A Program for Missing Data", developed by Honaker, King and Blackwell is available as R-package.
- Missing data imputation can easily be handled in Bayesian methods.
- Amelia provides a Bayesian setup to handle missing data imputation

Amelia : The R package for Missing Data Imputation

- "AMELIA II: A Program for Missing Data", developed by Honaker, King and Blackwell is available as R-package.
- Missing data imputation can easily be handled in Bayesian methods.
- Amelia provides a Bayesian setup to handle missing data imputation

- Observation-level priors:

    - If one has an additional prior information about missing data values based on previous research, academic consensus, or personal experience. Amelia can incorporate this information to produce vastly improved imputations.

    - The Amelia algorithm allows to include informative Bayesian priors about individual missing data cells.

- Observation-level priors:
- If one has an additional prior information about missing data values based on previous research, academic consensus, or personal experience. Amelia can incorporate this information to produce vastly improved imputations.
- The Amelia algorithm allows to include informative Bayesian priors about individual missing data cells.

Amelia : The R package for Missing Data Imputation

- Observation-level priors:
- If one has an additional prior information about missing data values based on previous research, academic consensus, or personal experience. Amelia can incorporate this information to produce vastly improved imputations.
- The Amelia algorithm allows to include informative Bayesian priors about individual missing data cells.

- The incorporation of priors follows basic Bayesian analysis where the imputation turns out to be a weighted average of the model-based imputation.
- The prior mean, where the weights are functions of the relative strength of the data and prior.
- If the model predicts well, the imputation will down-weight the prior, and vice versa (Honaker and King, 2010).

- The incorporation of priors follows basic Bayesian analysis where the imputation turns out to be a weighted average of the model-based imputation.
- The prior mean, where the weights are functions of the relative strength of the data and prior.
- If the model predicts well, the imputation will down-weight the prior, and vice versa (Honaker and King, 2010).

- The incorporation of priors follows basic Bayesian analysis where the imputation turns out to be a weighted average of the model-based imputation.

- The prior mean, where the weights are functions of the relative strength of the data and prior.

- If the model predicts well, the imputation will down-weight the prior, and vice versa (Honaker and King, 2010).

Application

- we consider mtcars data available R
- We consider a subset data with three variables in our analysis (i) mpg, (ii) hp and (iii) disp
- We randomly delete certain cells of the data
- We impute the missing values using Amelia and then we cross-check against the actual cell value

Application

- we consider mtcars data available R
- We consider a subset data with three variables in our analysis (i) mpg, (ii) hp and (iii) disp
- We randomly delete certain cells of the data
- We impute the missing values using Amelia and then we cross-check against the actual cell value

Application

- we consider mtcars data available R
- We consider a subset data with three variables in our analysis (i) mpg, (ii) hp and (iii) disp
- We randomly delete certain cells of the data
- We impute the missing values using Amelia and then we cross-check against the actual cell value

- we consider mtcars data available R
- We consider a subset data with three variables in our analysis (i) mpg, (ii) hp and (iii) disp
- We randomly delete certain cells of the data
- We impute the missing values using Amelia and then we cross-check against the actual cell value

Amelia : The R package for Missing Data Imputation

```
-- Imputation 1 --

  1  2  3  4  5  6  7  8  9 10 11 12 13 14

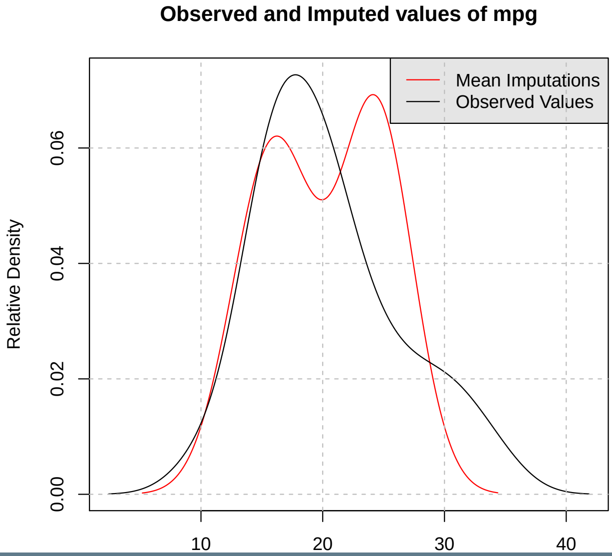
-- Imputation 2 --

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15

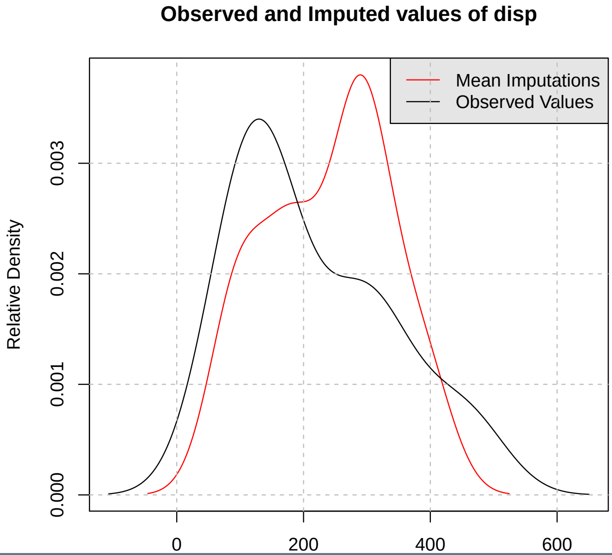
-- Imputation 3 --

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
```

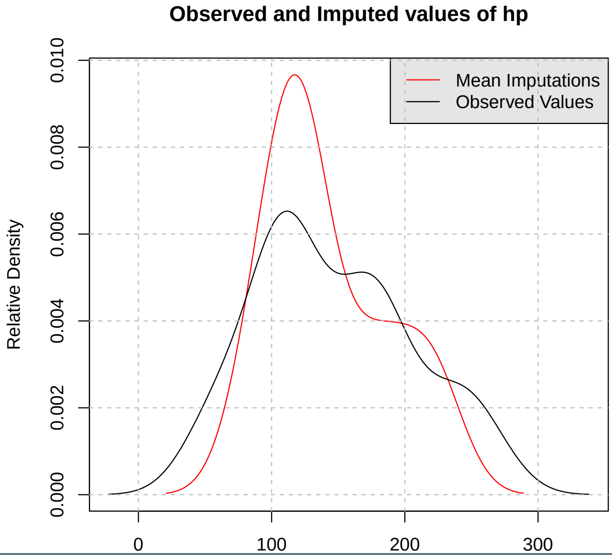
Amelia : The R package for Missing Data Imputation



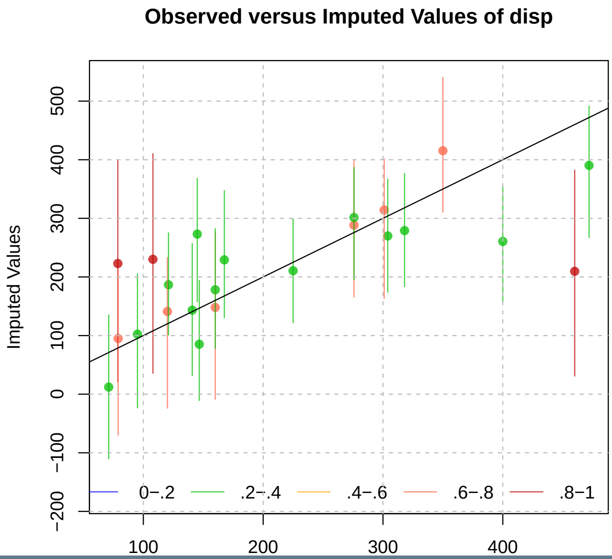
Amelia : The R package for Missing Data Imputation



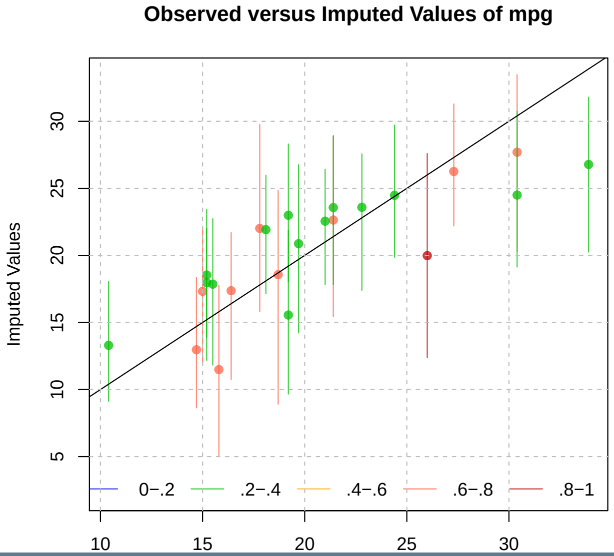
Amelia : The R package for Missing Data Imputation



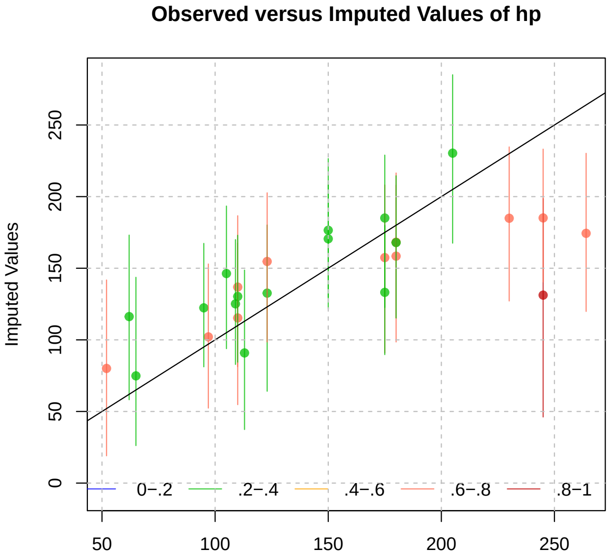
Amelia : The R package for Missing Data Imputation



Amelia : The R package for Missing Data Imputation



Amelia : The R package for Missing Data Imputation



A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

MHRD Govt. of India

Subject: Statistics

Paper: Statistical Inference

Module: Nonparametric Bayesian Analysis: Dirichlet Process Models

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Introduction to Dirichlet Process models

2. Stick Breaking and Chinese Resturant Process

3. Gibbs Sampler for Dirichlet Process models

4. Application

Outline

1. Introduction to Dirichlet Process models
2. Stick Breaking and Chinese Resturant Process
3. Gibbs Sampler for Dirichlet Process models
4. Application

Outline

1. Introduction to Dirichlet Process models
2. Stick Breaking and Chinese Resturant Process
3. Gibbs Sampler for Dirichlet Process models
4. Application

Outline

1. Introduction to Dirichlet Process models

2. Stick Breaking and Chinese Resturant Process

3. Gibbs Sampler for Dirichlet Process models

4. Application

- The Dirichlet process is an infinite-dimensional generalization of the Dirichlet distribution.
- It can be used to set as prior on unknown distributions.
- These unknown densities can be used to extend finite component mixture models to infinite component mixture models.

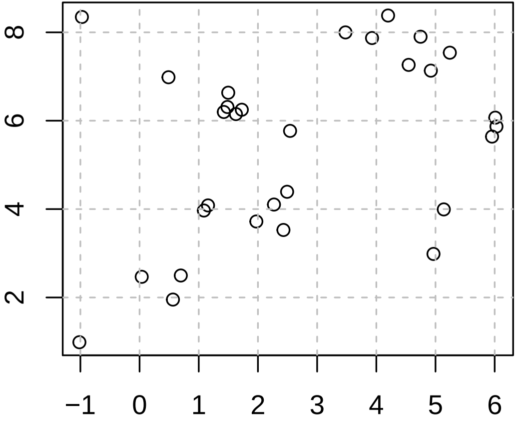
- The Dirichlet process is an infinite-dimensional generalization of the Dirichlet distribution.
- It can be used to set as prior on unknown distributions.
- These unknown densities can be used to extend finite component mixture models to infinite component mixture models.

Introduction

- The Dirichlet process is an infinite-dimensional generalization of the Dirichlet distribution.
- It can be used to set as prior on unknown distributions.
- These unknown densities can be used to extend finite component mixture models to infinite component mixture models.

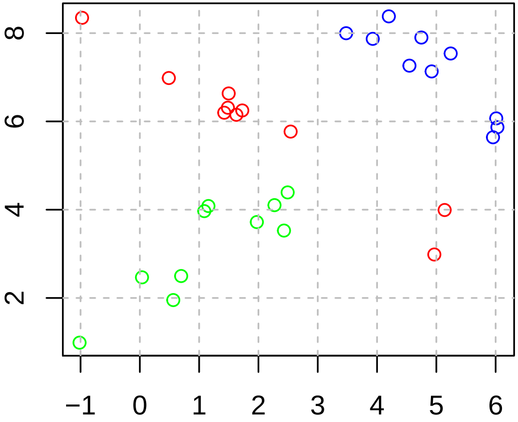
Motivation

- We are given a data set as follows:

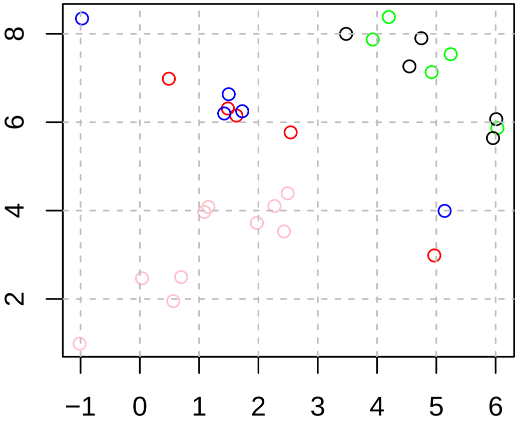


Motivation

- Even if we know the data is from a mixture of Gaussian



- it is difficult to tell how many distribution is actually there.



- Let $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$

$$\Theta \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_m)$$

$$P(\theta_1, \theta_2, ..., \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.
- Beta distribution is the special case of a Dirichlet for 2 dimensions.

Dirichlet Distribution

- Let $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$

  $$\Theta \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_m)$$

  $$P(\theta_1, \theta_2, ..., \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.

- Beta distribution is the special case of a Dirichlet for 2 dimensions.

Dirichlet Distribution

- Let $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$

  $$\Theta \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_m)$$

  $$P(\theta_1, \theta_2, ..., \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.

- Beta distribution is the special case of a Dirichlet for 2 dimensions.

- Let $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$

  $\Theta \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_m)$

$$P(\theta_1, \theta_2, ..., \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.

- Beta distribution is the special case of a Dirichlet for 2 dimensions.

Bayesian Histogram

- $y_i \overset{iid}{\sim} f()$, where density $f()$ is unknown
- The goal is to obtain a Bayes estimate of the density f.
- The histogram is often used as a simple form of density estimate.

- $y_i \overset{iid}{\sim} f()$, where density $f()$ is unknown
- The goal is to obtain a Bayes estimate of the density f.
- The histogram is often used as a simple form of density estimate.

Bayesian Histogram

- $y_i \overset{iid}{\sim} f()$, where density $f()$ is unknown
- The goal is to obtain a Bayes estimate of the density f.
- The histogram is often used as a simple form of density estimate.

Bayesian Histogram

- Assume we have prespecified knots $\xi = (\xi_0, \xi_1, ..., \xi_k)$ to define our histogram estimate with $\xi_0 < \xi_1 < ... < \xi_k$ and $y_i \in [\xi_0, \xi_k]$.

- A probability model for the density that is analogous to the histogram is as follows:

$$f(y) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_k} \frac{\pi_h}{(\xi_h - \xi_{h-1})}, \quad y \in \mathcal{R}$$

with $\pi = (\pi_1, ..., \pi_k)$ an unknown probability vector.

- Assume we have prespecified knots $\xi = (\xi_0, \xi_1, ..., \xi_k)$ to define our histogram estimate with $\xi_0 < \xi_1 < ... < \xi_k$ and $y_i \in [\xi_0, \xi_k]$.

- A probability model for the density that is analogous to the histogram is as follows:

$$f(y) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_k} \frac{\pi_h}{(\xi_h - \xi_{h-1})}, \quad y \in \mathcal{R}$$

with $\pi = (\pi_1, ..., \pi_k)$ an unknown probability vector.

- Bayes specification with a prior distribution for the probabilities.

- Assume a Dirichlet$(a_1, ..., a_k)$ prior distribution for π,

$$p(\pi|a) = \frac{\prod_{h=1}^{k} \Gamma(a)}{\Gamma(\sum_{h=1}^{k} a_h)} \prod_{h=1}^{k} \pi_h^{a_h - 1}$$

- The posterior distribution of π is

$$p(\pi|y, a) \propto \prod_{h=1}^{k} \pi_h^{a_h - 1} \prod_{i:y_i \in (\xi_{h-1}, \xi_h]} \frac{\pi_h}{\xi_h - \xi_{h-1}}$$

Bayesian Histogram

- Bayes specification with a prior distribution for the probabilities.

- Assume a Dirichlet$(a_1, ..., a_k)$ prior distribution for π,

$$p(\pi|a) = \frac{\prod_{h=1}^{k} \Gamma(a)}{\Gamma(\sum_{h=1}^{k} a_h)} \prod_{h=1}^{k} \pi_h^{a_h - 1}$$

- The posterior distribution of π is

$$p(\pi|y, a) \propto \prod_{h=1}^{k} \pi_h^{a_h - 1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h]} \frac{\pi_h}{\xi_h - \xi_{h-1}}$$

- Bayes specification with a prior distribution for the probabilities.

- Assume a Dirichlet$(a_1, ..., a_k)$ prior distribution for π,

$$p(\pi|a) = \frac{\prod_{h=1}^{k} \Gamma(a)}{\Gamma(\sum_{h=1}^{k} a_h)} \prod_{h=1}^{k} \pi_h^{a_h - 1}$$

- The posterior distribution of π is

$$p(\pi|y, a) \propto \prod_{h=1}^{k} \pi_h^{a_h - 1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h]} \frac{\pi_h}{\xi_h - \xi_{h-1}}$$

- The posterior distribution of π is

$$\begin{aligned} p(\pi|y,a) &\propto \prod_{h=1}^{k} \pi_h^{a_h-1} \prod_{i:y_i\in(\xi_{h-1},\xi_h]} \frac{\pi_h}{\xi_h-\xi_{h-1}} \\ &\propto \prod_{h=1}^{k} \pi_h^{a_h+n_h-1} \stackrel{D}{=} \text{Dirichlet}(a_1+n_1,...,a_k+n_k) \end{aligned}$$

where $n_h = \sum_i 1_{\xi_{h-1}<y_i\leq\xi_h}$ is the number of observations falling in the h^{th} histogram bin.

- The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots.

- The posterior distribution of π is

$$p(\pi|y,a) \propto \prod_{h=1}^{k} \pi_h^{a_h-1} \prod_{i:y_i\in(\xi_{h-1},\xi_h]} \frac{\pi_h}{\xi_h-\xi_{h-1}}$$

$$\propto \prod_{h=1}^{k} \pi_h^{a_h+n_h-1} \stackrel{D}{=} \text{Dirichlet}(a_1+n_1,...,a_k+n_k)$$

where $n_h = \sum_i 1_{\xi_{h-1}<y_i\leq\xi_h}$ is the number of observations falling in the h^{th} histogram bin.

- The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots.

- The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots.

- This approach allows prior information to be included and allows easy production of interval estimates, and hence has some practical advantages over classical histogram estimators.

- The Dirichlet prior distribution is perhaps not the best choice due to the lack of smoothing across adjacent bins, but it does have the advantage of conjugacy and simplicity in interpretation of the hyperparameters.

- The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots.
- This approach allows prior information to be included and allows easy production of interval estimates, and hence has some practical advantages over classical histogram estimators.
- The Dirichlet prior distribution is perhaps not the best choice due to the lack of smoothing across adjacent bins, but it does have the advantage of conjugacy and simplicity in interpretation of the hyperparameters.

- The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots.

- This approach allows prior information to be included and allows easy production of interval estimates, and hence has some practical advantages over classical histogram estimators.

- The Dirichlet prior distribution is perhaps not the best choice due to the lack of smoothing across adjacent bins, but it does have the advantage of conjugacy and simplicity in interpretation of the hyperparameters.

- A *Dirichlet Process* is a distribution over distributions.
- Let G be Dirichlet Process distributed:

$$G \sim P(\alpha, G_0)$$

  1. G_0 is a base distribution
  2. α is a positive scaling parameter
- G is a random probability measure that has the same support as G_0.

- A *Dirichlet Process* is a distribution over distributions.
- Let G be Dirichlet Process distributed:

$$G \sim P(\alpha, G_0)$$

  1. G_0 is a base distribution
  2. α is a positive scaling parameter
- G is a random probability measure that has the same support as G_0.

- A *Dirichlet Process* is a distribution over distributions.
- Let G be Dirichlet Process distributed:

$$G \sim P(\alpha, G_0)$$

1. G_0 is a base distribution
2. α is a positive scaling parameter

- G is a random probability measure that has the same support as G_0.

Dirichlet Process

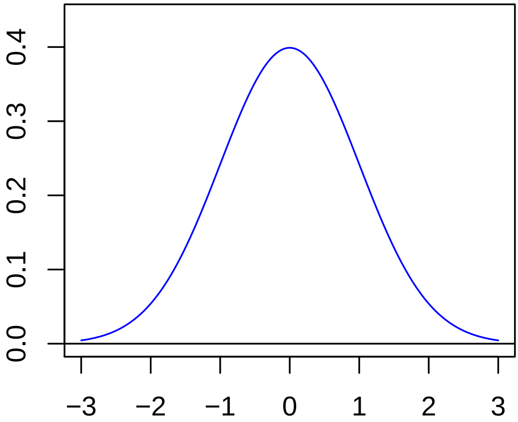
- A *Dirichlet Process* is a distribution over distributions.
- Let G be Dirichlet Process distributed:

$$G \sim P(\alpha, G_0)$$

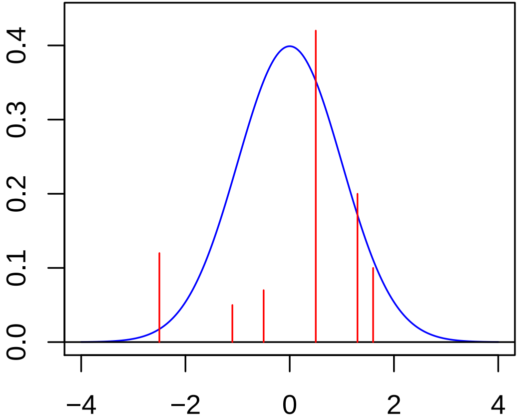
1. G_0 is a base distribution
2. α is a positive scaling parameter

- G is a random probability measure that has the same support as G_0.

- Consider Gaussian G_0



- $G \sim DP(\alpha, G_0)$



- G_0 is continuous, so the probability that any two samples are same is zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses
- Hence, there is a non-zero probability of two samples colliding

- G_0 is continuous, so the probability that any two samples are same is zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses
- Hence, there is a non-zero probability of two samples colliding

- G_0 is continuous, so the probability that any two samples are same is zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses
- Hence, there is a non-zero probability of two samples colliding

Sample from Dirichlet Processes

- $G \sim DP(\alpha, G_0)$
- $X_i|G \sim G$ for $i = \{1, 2, ..., n\}$ (iid given G)
- Marginalizing out G introduces dependencies between X_i

$$P(X_1, X_2, ..., X_n) = \int P(G) \prod_{i=1}^{n} P(X_i|G) dG$$

Sample from Dirichlet Processes

- $G \sim DP(\alpha, G_0)$
- $X_i|G \sim G$ for $i = \{1, 2, ..., n\}$ (iid given G)
- Marginalizing out G introduces dependencies between X_i

$$P(X_1, X_2, ..., X_n) = \int P(G) \prod_{i=1}^{n} P(X_i|G) dG$$

Sample from Dirichlet Processes

- $G \sim DP(\alpha, G_0)$
- $X_i|G \sim G$ for $i = \{1, 2, ..., n\}$ (iid given G)
- Marginalizing out G introduces dependencies between X_i

$$P(X_1, X_2, ..., X_n) = \int P(G) \prod_{i=1}^{n} P(X_i|G) dG$$

Sample from Dirichlet Processes

- $G \sim DP(\alpha, G_0)$
- $X_i|G \sim G$ for $i = \{1, 2, ..., n\}$ (iid given G)
- Marginalizing out G introduces dependencies between X_i

$$P(X_1, X_2, ..., X_n) = \int P(G) \prod_{i=1}^{n} P(X_i|G) dG$$

- Assume we view these variables in a specific order and are interested in the behavior of X_n given the previous $n-1$ observations.

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

- Assume we view these variables in a specific order and are interested in the behavior of X_n given the previous $n-1$ observations.

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

- Assume we view these variables in a specific order and are interested in the behavior of X_n given the previous $n-1$ observations.

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

Samples from a Dirichlet Process

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$\begin{aligned} P(X_1, X_2, ..., X_n) &= \int P(G) \prod_{i=1}^{n} P(X_i | G) dG \\ &= \frac{\alpha^K \prod_{k=1}^{K} (\#(X_k^*) - 1)!}{\alpha(\alpha+1)...(\alpha+n-1)} \prod_{k=1}^{K} G_0(X_k^*) \end{aligned}$$

- Notice that the above formulation of the joint distribution does not depend on the order we consider the variables.

Samples from a Dirichlet Process

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$\begin{aligned} P(X_1, X_2, ..., X_n) &= \int P(G) \prod_{i=1}^{n} P(X_i|G) dG \\ &= \frac{\alpha^K \prod_{k=1}^{K} (\#(X_k^*) - 1)!}{\alpha(\alpha+1)...(\alpha+n-1)} \prod_{k=1}^{K} G_0(X_k^*) \end{aligned}$$

- Notice that the above formulation of the joint distribution does not depend on the order we consider the variables.

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$\begin{aligned} P(X_1, X_2, ..., X_n) &= \int P(G) \prod_{i=1}^{n} P(X_i|G) dG \\ &= \frac{\alpha^K \prod_{k=1}^{K} (\#(X_k^*) - 1)!}{\alpha(\alpha+1)...(\alpha+n-1)} \prod_{k=1}^{K} G_0(X_k^*) \end{aligned}$$

- Notice that the above formulation of the joint distribution does not depend on the order we consider the variables.

Samples from a Dirichlet Process

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

Samples from a Dirichlet Process

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_i & \text{with prob} \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, 2, ..., K\}$$

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Blackwell-MacQueenUrn Scheme

- $G \sim DP(\alpha, G_0)$
- $X_n | G \sim G$
- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step n:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.
    Ref; Blackwell and MacQueen (1973)

Chinese Restaurant Process

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant.
- From the above conditional probability distribution, we can see that a new customer is more likely to sit at a table if there are already many people sitting there.
- However, with probability proportional to α, the customer will sit at a new table.
- This is known as "Clustering effect"

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant.
- From the above conditional probability distribution, we can see that a new customer is more likely to sit at a table if there are already many people sitting there.
- However, with probability proportional to α, the customer will sit at a new table.
- This is known as "Clustering effect"

$$X_n | X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant.
- From the above conditional probability distribution, we can see that a new customer is more likely to sit at a table if there are already many people sitting there.
- However, with probability proportional to α, the customer will sit at a new table.
- This is known as "Clustering effect"

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant.
- From the above conditional probability distribution, we can see that a new customer is more likely to sit at a table if there are already many people sitting there.
- However, with probability proportional to α, the customer will sit at a new table.
- This is known as "Clustering effect"

$$X_n|X_1, X_2, ..., X_{n-1} = \begin{cases} X_k^* & \text{with prob} \frac{\#_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob} \frac{\alpha}{n-1+\alpha} \end{cases}$$

- Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant.
- From the above conditional probability distribution, we can see that a new customer is more likely to sit at a table if there are already many people sitting there.
- However, with probability proportional to α, the customer will sit at a new table.
- This is known as "Clustering effect"

Stick Breaking

- So far, we have discussed properties of a distribution G drawn from a Dirichlet Process
- In 1994, Sethuraman developed a constructive way of forming G, known as - "stick breaking"
- "Stick breaking" helped practitioner to implement the Dirichlet Process in a real way.

Stick Breaking

- So far, we have discussed properties of a distribution G drawn from a Dirichlet Process
- In 1994, Sethuraman developed a constructive way of forming G, known as - "stick breaking"
- "Stick breaking" helped practitioner to implement the Dirichlet Process in a real way.

Stick Breaking

- So far, we have discussed properties of a distribution G drawn from a Dirichlet Process
- In 1994, Sethuraman developed a constructive way of forming G, known as - "stick breaking"
- "Stick breaking" helped practitioner to implement the Dirichlet Process in a real way.

Stick Breaking

- $V_1, V_2, ..., V_i, ... \sim Beta(1, \alpha)$
- $f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha - 1}$
- $X_1^*, X_2^*, ..., X_i^*, ... \sim G_0$
- $\pi_i(v) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- $G = \sum_{i=1}^{\infty} \pi_i(v)\delta_{X_i^*}$

Stick Breaking

- $V_1, V_2, ..., V_i, ... \sim Beta(1, \alpha)$
- $f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$
- $X_1^*, X_2^*, ..., X_i^*, ... \sim G_0$
- $\pi_i(v) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- $G = \sum_{i=1}^{\infty} \pi_i(v)\delta_{X_i^*}$

Stick Breaking

- $V_1, V_2, ..., V_i, ... \sim Beta(1, \alpha)$
- $f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$
- $X_1^*, X_2^*, ..., X_i^*, ... \sim G_0$
- $\pi_i(v) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- $G = \sum_{i=1}^{\infty} \pi_i(v)\delta_{X_i^*}$

Stick Breaking

- $V_1, V_2, ..., V_i, ... \sim Beta(1, \alpha)$
- $f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$
- $X_1^*, X_2^*, ..., X_i^*, ... \sim G_0$
- $\pi_i(v) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- $G = \sum_{i=1}^{\infty} \pi_i(v)\delta_{X_i^*}$

- $V_1, V_2, ..., V_i, ... \sim Beta(1, \alpha)$
- $f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$
- $X_1^*, X_2^*, ..., X_i^*, ... \sim G_0$
- $\pi_i(v) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- $G = \sum_{i=1}^{\infty} \pi_i(v)\delta_{X_i^*}$

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  $\ldots$

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  ...

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  $\ldots$

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  ...

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  ...

Stick Breaking

- Draw X_1^* from G_0
- Draw v_1 from $Beta(1, \alpha)$
- $\pi_1 = v_1$
- Draw X_2^* from G_0
- Draw v_2 from $Beta(1, \alpha)$
- $\pi_2 = v_2(1 - v_1)$

  ...

Stick Breaking

- Let α be a positive, real-valued scalar
- Let G_0 be a non-atomic probability distribution over support set A
- If $G \sim DP(\alpha, G_0)$, then for any finite set of partitions A:

$$A_1 \cup A_2 \cup ... \cup A_k$$

then

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha G_0(A_1), ..., G_0(A_k))$$

Stick Breaking

- Let α be a positive, real-valued scalar
- Let G_0 be a non-atomic probability distribution over support set A
- If $G \sim DP(\alpha, G_0)$, then for any finite set of partitions A:

$$A_1 \cup A_2 \cup ... \cup A_k$$

then

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha G_0(A_1), ..., G_0(A_k))$$

- Let α be a positive, real-valued scalar
- Let G_0 be a non-atomic probability distribution over support set A
- If $G \sim DP(\alpha, G_0)$, then for any finite set of partitions A:

$$A_1 \cup A_2 \cup ... \cup A_k$$

then

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha G_0(A_1), ..., G_0(A_k))$$

DP Mixture Model

- general kernel mixture model

$$f(y|P) = \int (y|\theta) dP(\theta)$$

where $(|\theta)$ is a kernel with θ include location and scale parameters

- a DP prior on P leads to

$$f(y) = \sum_{h=1}^{\infty} \pi_h (y|\theta_h^*)$$

where π denote the probability weights are sampled from a DP stick-breaking process with parameter α, and with $\theta_h \sim P_0$ independently for $h = 1, 2, ...\infty$.

DP Mixture Model

- general kernel mixture model

$$f(y|P) = \int (y|\theta) dP(\theta)$$

where $(|\theta)$ is a kernel with θ include location and scale parameters

- a DP prior on P leads to

$$f(y) = \sum_{h=1}^{\infty} \pi_h (y|\theta_h^*)$$

where π denote the probability weights are sampled from a DP stick-breaking process with parameter α, and with $\theta_h \sim P_0$ independently for $h = 1, 2, ...\infty$.

- A key question is how to conduct posterior computation under the DP?
- This initially seems problematic in that the mixing measure P is characterized by infinitely many parameters
- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$

- A key question is how to conduct posterior computation under the DP?
- This initially seems problematic in that the mixing measure P is characterized by infinitely many parameters
- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$

- A key question is how to conduct posterior computation under the DP?

- This initially seems problematic in that the mixing measure P is characterized by infinitely many parameters

- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$

Polya Urn Scheme

- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$.

- In particular, marginalizing out P, we obtain the *Polya urn* predictive rule,

$$p(\theta_i|\theta_1, ..., \theta_i) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{\theta_j}$$

- Chinese restaurant process metaphor is commonly used in describing the Polya urn scheme.

- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$.

- In particular, marginalizing out P, we obtain the *Polya urn* predictive rule,

$$p(\theta_i|\theta_1, ..., \theta_i) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{\theta_j}$$

- Chinese restaurant process metaphor is commonly used in describing the Polya urn scheme.

- A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta = (\theta_1, ..., \theta_n)$.

- In particular, marginalizing out P, we obtain the *Polya urn* predictive rule,

$$p(\theta_i|\theta_1, ..., \theta_i) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{\theta_j}$$

- Chinese restaurant process metaphor is commonly used in describing the Polya urn scheme.

Blocked Gibbs Sampler

- By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P.

- By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically.

- One approach for avoiding marginalization is to rely on the construction

$$f(y) = \sum_{h=1}^{\infty} \pi_h (y|\theta_h^*)$$

- Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero.

- By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P.

- By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically.

- One approach for avoiding marginalization is to rely on the construction

$$f(y) = \sum_{h=1}^{\infty} \pi_h(y|\theta_h^*)$$

- Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero.

Blocked Gibbs Sampler

- By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P.

- By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically.

- One approach for avoiding marginalization is to rely on the construction

$$f(y) = \sum_{h=1}^{\infty} \pi_h (y|\theta_h^*)$$

- Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero.

Blocked Gibbs Sampler

- By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P.

- By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically.

- One approach for avoiding marginalization is to rely on the construction

$$f(y) = \sum_{h=1}^{\infty} \pi_h(y|\theta_h^*)$$

- Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero.

Blocked Gibbs Sampler

- By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P.

- By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically.

- One approach for avoiding marginalization is to rely on the construction

$$f(y) = \sum_{h=1}^{\infty} \pi_h(y|\theta_h^*)$$

- Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero.

Blocked Gibbs Sampler

- Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1 Update $S_i \in \{1, 2, ..., N\}$ by multinomial sampling with

$$Pr(S_i = h|-) = \frac{\pi_h(y_i|\theta_h^*)}{\sum_{h'=1}^{N} \pi_{h'}(y_i|\theta_{h'}^*)}$$

where $S_i = h$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2 Update the stick-breaking weight V_h, $h = 1, 2, ..., N-1$ from $Beta(1 + n_h, \alpha + \sum_{h'=h+1}^{N} n_{h'})$.

3 Update θ_h^* , $h = 1, 2, ..., N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_h = 0$ sampled from the prior P_0.

- This algorithm involves simple sampling steps and is straightforward to implement.

- Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1. Update $S_i \in \{1, 2, ..., N\}$ by multinomial sampling with

$$Pr(S_i = h|-) = \frac{\pi_h(y_i|\theta_h^*)}{\sum_{h'=1}^{N} \pi_{h'}(y_i|\theta_{h'}^*)}$$

   where $S_i = h$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2. Update the stick-breaking weight V_h, $h = 1, 2, ..., N-1$ from $Beta(1 + n_h, \alpha + \sum_{h'=h+1}^{N} n_{h'})$.
3. Update θ_h^* , $h = 1, 2, ..., N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_h = 0$ sampled from the prior P_0.

- This algorithm involves simple sampling steps and is straightforward to implement.

Blocked Gibbs Sampler

- Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1 Update $S_i \in \{1, 2, ..., N\}$ by multinomial sampling with

$$Pr(S_i = h|-) = \frac{\pi_h(y_i|\theta_h^*)}{\sum_{h'=1}^{N} \pi_{h'}(y_i|\theta_{h'}^*)}$$

where $S_i = h$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2 Update the stick-breaking weight V_h, $h = 1, 2, ..., N-1$ from $Beta(1 + n_h, \alpha + \sum_{h'=h+1}^{N} n_{h'})$.

3 Update θ_h^*, $h = 1, 2, ..., N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_h = 0$ sampled from the prior P_0.

- This algorithm involves simple sampling steps and is straightforward to implement.

- Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1 Update $S_i \in \{1, 2, ..., N\}$ by multinomial sampling with

$$Pr(S_i = h|-) = \frac{\pi_h(y_i|\theta_h^*)}{\sum_{h'=1}^{N} \pi_{h'}(y_i|\theta_{h'}^*)}$$

where $S_i = h$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2 Update the stick-breaking weight V_h, $h = 1, 2, ..., N-1$ from $Beta(1 + n_h, \alpha + \sum_{h'=h+1}^{N} n_{h'})$.

3 Update θ_h^* , $h = 1, 2, ..., N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_h = 0$ sampled from the prior P_0.

- This algorithm involves simple sampling steps and is straightforward to implement.

Blocked Gibbs Sampler

- Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1. Update $S_i \in \{1, 2, ..., N\}$ by multinomial sampling with

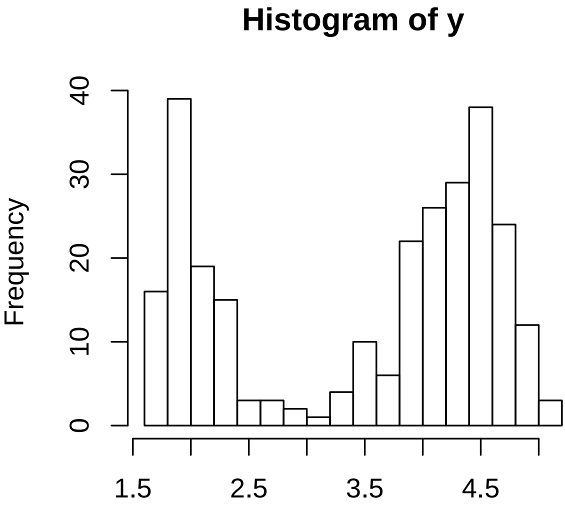
$$Pr(S_i = h|-) = \frac{\pi_h(y_i|\theta_h^*)}{\sum_{h'=1}^{N} \pi_{h'}(y_i|\theta_{h'}^*)}$$

   where $S_i = h$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2. Update the stick-breaking weight V_h, $h = 1, 2, ..., N-1$ from $Beta(1 + n_h, \alpha + \sum_{h'=h+1}^{N} n_{h'})$.
3. Update θ_h^* , $h = 1, 2, ..., N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_h = 0$ sampled from the prior P_0.

- This algorithm involves simple sampling steps and is straightforward to implement.

- Consiter faithful dataset



Application

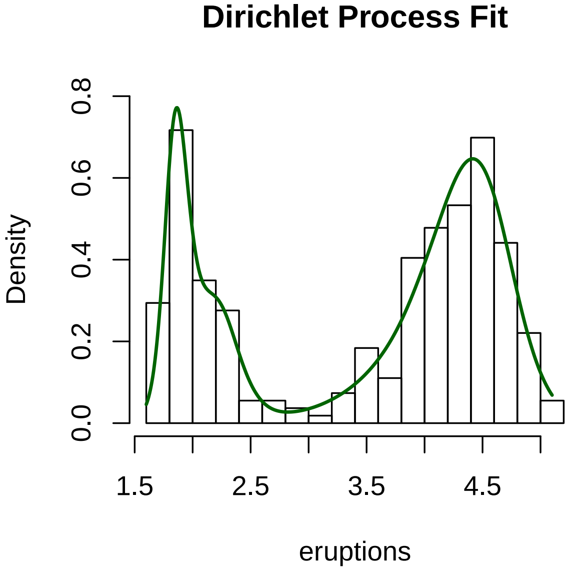
- Consiter $\mathrm{faithful}$ dataset
- Clearly it is a bimodal data.

Application

- Consiter faithful dataset
- Clearly it is a bimodal data.

Application

- Consiter faithful dataset
- Clearly it is a bimodal data.


Dirichlet Process Fit
Density
0.0
0.2
0.4
0.6
0.8
1.5
2.5
3.5
4.5
eruptions

A Gateway to all Postgraduate Courses

An MHRD project under its National Mission on Education through ICT (NME-ICT)

MHRD Govt. of India

सत्यमेव जयते

Subject: Statistics

Paper: Statistical Inference

Module: Gaussian Process Prior for Non-Parametric Regression

Development Team

Principal investigator: *Dr. Bhaswati Ganguli, Professor, Department of Statistics, University of Calcutta*

Paper co-ordinator: *Dr. Dipak K Dey, Associate Dean and BOT Distinguished Professor, Department of Statistics, University of Connecticut*

Content writer: *Dr. Sourish Das, Assistant Professor, Chennai Mathematical Institute*

Content reviewer: *Department of Statistics, University of Calcutta*

Outline

1. Introduction to Gaussian Process
2. Function-Space View
3. Application

Outline

1. Introduction to Gaussian Process
2. Function-Space View
3. Application

Outline

1. Introduction to Gaussian Process
2. Function-Space View
3. Application

- The Gaussian process is an infinite-dimensional generalization of the Gaussian distribution.

- It can be used to set as prior over unknown function.

- In this module, we describe Gaussian process methods for regression problems. For more detail see Rasmussen and Williams (2006).

Introduction

- The Gaussian process is an infinite-dimensional generalization of the Gaussian distribution.
- It can be used to set as prior over unknown function.
- In this module, we describe Gaussian process methods for regression problems. For more detail see Rasmussen and Williams (2006).

- The Gaussian process is an infinite-dimensional generalization of the Gaussian distribution.

- It can be used to set as prior over unknown function.

- In this module, we describe Gaussian process methods for regression problems. For more detail see Rasmussen and Williams (2006).

Introduction

- There are several ways to interpret Gaussian process (GP) regression models.
- One can think of a Gaussian process as defining a distribution over functions, and inference taking place directly in the space of functions, *the function-space* view.
- An alternate equivalent view is *weight-space view*.

- There are several ways to interpret Gaussian process (GP) regression models.
- One can think of a Gaussian process as defining a distribution over functions, and inference taking place directly in the space of functions, *the function-space* view.
- An alternate equivalent view is *weight-space view*.

- There are several ways to interpret Gaussian process (GP) regression models.
- One can think of a Gaussian process as defining a distribution over functions, and inference taking place directly in the space of functions, *the function-space* view.
- An alternate equivalent view is *weight-space view*.

Weight-space view

- We have a training set $\mathcal{D} = \{(x_i, y_i) | i = 1, 2, ..., n\}$, where $\mathbf{x}$ denotes an input vector (covariates) of dimension D.
- y denotes a scalar output or target (dependent variable);
- the column vector inputs for all n cases are aggregated in the $D \times n$ design matrix X and the targets are collected in the vector y
- so we can write $\mathcal{D} = \{X, y\}$

Weight-space view

- We have a training set $\mathcal{D} = \{(x_i, y_i) | i = 1, 2, ..., n\}$, where $\mathbf{x}$ denotes an input vector (covariates) of dimension D.
- y denotes a scalar output or target (dependent variable);
- the column vector inputs for all n cases are aggregated in the $D \times n$ design matrix X and the targets are collected in the vector y
- so we can write $\mathcal{D} = \{X, y\}$

- We have a training set $\mathcal{D} = \{(x_i, y_i) | i = 1, 2, ..., n\}$, where $\mathbf{x}$ denotes an input vector (covariates) of dimension D.
- y denotes a scalar output or target (dependent variable);
- the column vector inputs for all n cases are aggregated in the $D \times n$ design matrix X and the targets are collected in the vector y
- so we can write $\mathcal{D} = \{X, y\}$

- We have a training set $\mathcal{D} = \{(x_i, y_i) | i = 1, 2, ..., n\}$, where $\mathbf{x}$ denotes an input vector (covariates) of dimension D.
- y denotes a scalar output or target (dependent variable);
- the column vector inputs for all n cases are aggregated in the $D \times n$ design matrix X and the targets are collected in the vector y
- so we can write $\mathcal{D} = \{X, y\}$

Weight-Space View

- standard linear regression model with Gaussian noise

$$f(x) = x^T \omega, \quad y = f(x) + \epsilon,$$

where x is the input vector, ω is a vector of weights (parameters) of the linear models, f is the function value and y is the observed target value

$$\epsilon \sim N(0, \sigma_n^2)$$

- This noise assumption together with the model directly gives rise to the likelihood.

Weight-Space View

- standard linear regression model with Gaussian noise

$$f(x) = x^T \omega, \quad y = f(x) + \epsilon,$$

where x is the input vector, ω is a vector of weights (parameters) of the linear models, f is the function value and y is the observed target value

$$\epsilon \sim N(0, \sigma_n^2)$$

- This noise assumption together with the model directly gives rise to the likelihood.

- standard linear regression model with Gaussian noise

$$f(x) = x^T\omega, \quad y = f(x) + \epsilon,$$

where x is the input vector, ω is a vector of weights (parameters) of the linear models, f is the function value and y is the observed target value

$$\epsilon \sim N(0, \sigma_n^2)$$

- This noise assumption together with the model directly gives rise to the likelihood.

- The probability density of the observations given the parameters, which is factored over cases in the training set (because of the independence assumption) to give

$$
\begin{aligned}
p(y|X,\omega) &= \prod_{i=1}^{n} p(y_i|x_i,\omega) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left\{-\frac{(y_i - x_i^T\omega)^2}{2\sigma_n^2}\right\} \\
&= \frac{1}{2\pi\sigma_n^2}^{n/2} \exp\left\{-\frac{1}{2\sigma_n^2}|y - X^T\omega|^2\right\} \\
&= = \mathcal{N}(X^T\omega, \sigma_n^2\mathbf{I})
\end{aligned} \tag{1}
$$

Weight-Space View

- In the Bayesian setup we should specify a prior over the parameters, before we look at the observations.

- We put a zero mean Gaussian prior with covariance matrix Σ_p on the weights

$$\omega \sim \mathcal{N}(0, \Sigma_p).$$

- Inference in the Bayesian linear model is based on the posterior distribution posterior over the weights, computed by Bayes' rule

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$

$$p(\omega|y, X) = \frac{p(y|X, \omega)p(\omega)}{p(y|X)}$$

where $p(y|X) = \int p(y|X, \omega)p(\omega)d\omega$

Weight-Space View

- In the Bayesian setup we should specify a prior over the parameters, before we look at the observations.

- We put a zero mean Gaussian prior with covariance matrix Σ_p on the weights
$$\omega \sim \mathcal{N}(0, \Sigma_p).$$

- Inference in the Bayesian linear model is based on the posterior distribution posterior over the weights, computed by Bayes' rule
$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$
$$p(\omega|y, X) = \frac{p(y|X, \omega)p(\omega)}{p(y|X)}$$
where $p(y|X) = \int p(y|X, \omega)p(\omega)d\omega$

- In the Bayesian setup we should specify a prior over the parameters, before we look at the observations.

- We put a zero mean Gaussian prior with covariance matrix Σ_p on the weights
$$\omega \sim \mathcal{N}(0, \Sigma_p).$$

- Inference in the Bayesian linear model is based on the posterior distribution posterior over the weights, computed by Bayes' rule
$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$
$$p(\omega|y, X) = \frac{p(y|X, \omega)p(\omega)}{p(y|X)}$$
where $p(y|X) = \int p(y|X, \omega)p(\omega)d\omega$

Weight-Space View

- The form of the posterior distribution as Gaussian with mean $\bar{\omega}$ and covariance matrix A^{-1}

$$p(\omega|X, y) \sim \mathcal{N}(\bar{\omega}, A^{-1})$$

where $\bar{\omega} = \frac{1}{\sigma_n^2} A^{-1} X y$ and $A^{-1} = \sigma_n^{-2} X X^T + \Sigma_p^{-1}$.

Weight-Space View

- In a non-Bayesian setting the negative log prior is sometimes thought of as a penalty term, and the maximum a posteriori (MAP) estimate point is known as the penalized maximum likelihood estimate of the weights.
- The penalized maximum likelihood procedure is known in this case as ridge regression [Hoerl and Kennard, 1970] because of ridge regression the effect of the quadratic penalty term $\frac{1}{2}\omega^T\Sigma_p^{-1}\omega$ from log-prior

- In a non-Bayesian setting the negative log prior is sometimes thought of as a penalty term, and the maximum a posteriori (MAP) estimate point is known as the penalized maximum likelihood estimate of the weights.

- The penalized maximum likelihood procedure is known in this case as ridge regression [Hoerl and Kennard, 1970] because of ridge regression the effect of the quadratic penalty term $\frac{1}{2}\omega^T\Sigma_p^{-1}\omega$ from log-prior

- To make predictions for a test case we average over all possible parameter values, weighted by their posterior probability.
- Thus the predictive distribution for $f_* = f(x_*)$ at x_* is given by averaging the output of all possible linear models w.r.t. the Gaussian posterior

$$\begin{aligned} p(f_*|x_*, X, y) &= \int p(f_*|x_*, \omega)p(\omega|X, y)d\omega \\ &= \mathcal{N}\left(\frac{1}{\sigma_n^2}x_*^T A^{-1}Xy, x_*^T A^{-1}x_*\right) \end{aligned}$$

- To make predictions for a test case we average over all possible parameter values, weighted by their posterior probability.

- Thus the predictive distribution for $f_* = f(x_*)$ at x_* is given by averaging the output of all possible linear models w.r.t. the Gaussian posterior

$$
\begin{aligned}
p(f_*|x_*, X, y) &= \int p(f_*|x_*, \omega)p(\omega|X, y)d\omega \\
&= \mathcal{N}\left(\frac{1}{\sigma_n^2} x_*^T A^{-1} X y, x_*^T A^{-1} x_*\right)
\end{aligned}
$$

Projections of Inputs into Feature Space

- Bayesian linear model suffers from limited expressiveness.

- **simple idea** to overcome this problem is to first project the inputs into some high dimensional space using a set of basis feature space functions

- and then apply the linear model in this space instead of directly on the inputs themselves.

- For example, a scalar input x could be projected into the space of powers of x: $\phi(x) = (1, x, x^2, x^3, ...)$ to implement polynomial polynomial regression.

- Bayesian linear model suffers from limited expressiveness.
- **simple idea** to overcome this problem is to first project the inputs into some high dimensional space using a set of basis feature space functions
- and then apply the linear model in this space instead of directly on the inputs themselves.
- For example, a scalar input x could be projected into the space of powers of x: $\phi(x) = (1, x, x^2, x^3, ...)$ to implement polynomial polynomial regression.

Projections of Inputs into Feature Space

- Bayesian linear model suffers from limited expressiveness.
- **simple idea** to overcome this problem is to first project the inputs into some high dimensional space using a set of basis feature space functions
- and then apply the linear model in this space instead of directly on the inputs themselves.
- For example, a scalar input x could be projected into the space of powers of x: $\phi(x) = (1, x, x^2, x^3, ...)$ to implement polynomial polynomial regression.

Projections of Inputs into Feature Space

- Bayesian linear model suffers from limited expressiveness.

- **simple idea** to overcome this problem is to first project the inputs into some high dimensional space using a set of basis feature space functions

- and then apply the linear model in this space instead of directly on the inputs themselves.

- For example, a scalar input x could be projected into the space of powers of x: $\phi(x) = (1, x, x^2, x^3, ...)$ to implement polynomial polynomial regression.

Projections of Inputs into Feature Space

- As long as the projections are fixed functions (i.e. independent of the parameters w) the model is still linear in the parameters, and therefore analytically tractable

- Specifically, we introduce the function $\phi(x)$ which maps D-dimensional input vector x into an N dimensional feature space.

- Suppose $\Phi(X)$ be aggregation of column (x) for all cases in the training set. Now the model is

$$f(x) = \phi(x)^T \omega$$

where the vector of parameters now has length N

Projections of Inputs into Feature Space

- As long as the projections are fixed functions (i.e. independent of the parameters w) the model is still linear in the parameters, and therefore analytically tractable

- Specifically, we introduce the function $\phi(x)$ which maps D-dimensional input vector x into an N dimensional feature space.

- Suppose $\Phi(X)$ be aggregation of column (x) for all cases in the training set. Now the model is

$$f(x) = \phi(x)^T \omega$$

where the vector of parameters now has length N

- As long as the projections are fixed functions (i.e. independent of the parameters w) the model is still linear in the parameters, and therefore analytically tractable

- Specifically, we introduce the function $\phi(x)$ which maps D-dimensional input vector x into an N dimensional feature space.

- Suppose $\Phi(X)$ be aggregation of column (x) for all cases in the training set. Now the model is

$$f(x) = \phi(x)^T \omega$$

where the vector of parameters now has length N

Projections of Inputs into Feature Space

- The analysis for this model is analogous to the standard linear model, except that everywhere $\Phi(X)$ is substituted for X.

- Thus the predictive distribution becomes

$$f_*|x_*, X, y \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\phi(x_*)^T A^{-1}\Phi y, \phi(x_*)^T A^{-1}\phi(x_*)\right)$$

where $\Phi = \Phi(X)$ and $A = \sigma_n^{-2}\Phi\Phi^T + \Sigma_p^{-1}$

- To make predictions using this equation we need to invert the A matrix of size $N \times N$ which may not be convenient if N, the dimension of the feature space, is large.

Projections of Inputs into Feature Space

- The analysis for this model is analogous to the standard linear model, except that everywhere $\Phi(X)$ is substituted for X.

- Thus the predictive distribution becomes

$$f_*|x_*, X, y \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\phi(x_*)^T A^{-1}\Phi y, \phi(x_*)^T A^{-1}\phi(x_*)\right)$$

where $\Phi = \Phi(X)$ and $A = \sigma_n^{-2}\Phi\Phi^T + \Sigma_p^{-1}$

- To make predictions using this equation we need to invert the A matrix of size $N \times N$ which may not be convenient if N, the dimension of the feature space, is large.

Projections of Inputs into Feature Space

- The analysis for this model is analogous to the standard linear model, except that everywhere $\Phi(X)$ is substituted for X.

- Thus the predictive distribution becomes

$$f_*|x_*, X, y \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\phi(x_*)^T A^{-1}\Phi y, \phi(x_*)^T A^{-1}\phi(x_*)\right)$$

  where $\Phi = \Phi(X)$ and $A = \sigma_n^{-2}\Phi\Phi^T + \Sigma_p^{-1}$

- To make predictions using this equation we need to invert the A matrix of size $N \times N$ which may not be convenient if N, the dimension of the feature space, is large.

- We can rewrite the predictive distribution as

$$f_*|x_*, X, y \sim \mathcal{N}(\phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} y, \\ \phi_*^T \Sigma_{p*} - \phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} \Phi^T \Sigma_p \phi_*) \quad (2)$$

where $\phi(x_*) = \phi_*$ and defined $K = \Phi^T \Sigma_p \Phi$

- In equation (2) the feature space always enters in the form of $\phi_*^T \Sigma_p \Phi$, $\Phi^T \Sigma_p \Phi$, or $\phi_*^T{}_p \phi_*$

- The entries of these matrices are invariably of the form $\phi(x)^T \Sigma_p \phi(X')$

Projections of Inputs into Feature Space

- We can rewrite the predictive distribution as

$$\begin{aligned} f_*|x_*, X, y \quad \sim \quad & \mathcal{N}(\phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} y, \\ & \phi_*^T \Sigma_{p*} - \phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} \Phi^T \Sigma_p \phi_*) \end{aligned} \tag{2}$$

where $\phi(x_*) = \phi_*$ and defined $K = \Phi^T \Sigma_p \Phi$

- In equation (2) the feature space always enters in the form of $\phi_*^T \Sigma_p \Phi$, $\Phi^T \Sigma_p \Phi$, or $\phi_*^T {}_p \phi_*$

- The entries of these matrices are invariably of the form $\phi(x)^T \Sigma_p \phi(X')$

Projections of Inputs into Feature Space

- We can rewrite the predictive distribution as

$$\begin{aligned} f_*|x_*, X, y \quad \sim \quad & \mathcal{N}(\phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} y, \\ & \phi_*^T \Sigma_{p*} - \phi_*^T \Sigma_p \Phi (K + \sigma_n^2)^{-1} \Phi^T \Sigma_p \phi_*) \end{aligned} \tag{2}$$

where $\phi(x_*) = \phi_*$ and defined $K = \Phi^T \Sigma_p \Phi$

- In equation (2) the feature space always enters in the form of $\phi_*^T \Sigma_p \Phi$, $\Phi^T \Sigma_p \Phi$, or $\phi_*^T{}_p \phi_*$

- The entries of these matrices are invariably of the form $\phi(x)^T \Sigma_p \phi(X')$

- Define $k(x, x') = \phi(x)^T \Sigma_p \phi(x')$ known as a **covariance function** or **kernel function**.

- Notice tha $\phi(x)^T \Sigma_p \phi(x')$ is an inner product with respect to Σ_p.

- As Σ_p is positive definite we can define $\Psi(x) = \Sigma_p^{1/2} \phi(x)$ we obtain a simple dot product representation $k(x, x') = \Psi(x).\Psi(x')$.

- If an algorithm is defined solely in terms of inner products in input space then it can be lifted into feature space by replacing occurrences of those inner products by $k(x, x_0)$; this is sometimes called the **kernel trick**.

Projections of Inputs into Feature Space

- Define $k(x, x') = \phi(x)^T \Sigma_p \phi(x')$ known as a **covariance function** or **kernel function**.
- Notice tha $\phi(x)^T \Sigma_p \phi(x')$ is an inner product with respect to Σ_p.
- As Σ_p is positive definite we can define $\Psi(x) = \Sigma_p^{1/2} \phi(x)$ we obtain a simple dot product representation $k(x, x') = \Psi(x).\Psi(x')$.
- If an algorithm is defined solely in terms of inner products in input space then it can be lifted into feature space by replacing occurrences of those inner products by $k(x, x_0)$; this is sometimes called the **kernel trick**.

- Define $k(x, x') = \phi(x)^T \Sigma_p \phi(x')$ known as a **covariance function** or **kernel function**.
- Notice tha $\phi(x)^T \Sigma_p \phi(x')$ is an inner product with respect to Σ_p.
- As Σ_p is positive definite we can define $\Psi(x) = \Sigma_p^{1/2} \phi(x)$ we obtain a simple dot product representation $k(x, x') = \Psi(x).\Psi(x')$.
- If an algorithm is defined solely in terms of inner products in input space then it can be lifted into feature space by replacing occurrences of those inner products by $k(x, x_0)$; this is sometimes called the **kernel trick**.

- Define $k(x, x') = \phi(x)^T \Sigma_p \phi(x')$ known as a **covariance function** or **kernel function**.

- Notice tha $\phi(x)^T \Sigma_p \phi(x')$ is an inner product with respect to Σ_p.

- As Σ_p is positive definite we can define $\Psi(x) = \Sigma_p^{1/2} \phi(x)$ we obtain a simple dot product representation $k(x, x') = \Psi(x).\Psi(x')$.

- If an algorithm is defined solely in terms of inner products in input space then it can be lifted into feature space by replacing occurrences of those inner products by $k(x, x_0)$; this is sometimes called the **kernel trick**.

- **Definition**: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

- A Gaussian process is completely specified by its mean function and co- covariance and variance function. We define mean function $m(x)$ and covariance function $k(x, x')$ of real process $f(x)$ as

$$\begin{aligned} m(x) &= \mathrm{E}[f(x)] \\ k(x, x') &= \mathrm{E}[(f(x) - m(x))(f(x') - m(x'))] \end{aligned} \tag{3}$$

and will write the Gaussian process as

$$f(x) \sim \mathcal{GP}\left(m(x), k(x, x')\right) \tag{4}$$

Function-Space View

- **Definition**: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

- A Gaussian process is completely specified by its mean function and co- covariance and variance function. We define mean function $m(x)$ and covariance function $k(x, x')$ of real process $f(x)$ as

$$\begin{aligned} m(x) &= \mathrm{E}[f(x)] \\ k(x, x') &= \mathrm{E}[(f(x) - m(x))(f(x') - m(x'))] \end{aligned} \tag{3}$$

  and will write the Gaussian process as

$$f(x) \sim \mathcal{GP}\left(m(x), k(x, x')\right) \tag{4}$$

Function-Space View

- The random variables represent the value of the function $f(x)$ at location x.

- A Gaussian process is defined as a collection of random variables.

- The definition automatically implies a consistency requirement, which is also sometimes known as the marginalization property.

Function-Space View

- The random variables represent the value of the function $f(x)$ at location x.

- A Gaussian process is defined as a collection of random variables.

- The definition automatically implies a consistency requirement, which is also sometimes known as the marginalization property.

Function-Space View

- The random variables represent the value of the function $f(x)$ at location x.
- A Gaussian process is defined as a collection of random variables.
- The definition automatically implies a consistency requirement, which is also sometimes known as the marginalization property.

Function-Space View

- The marginalization property simply means that if the GP e.g. specifies $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$, then it must also specify $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ where Σ_{11} is the relevant submatrix of Σ

- A simple example of a Gaussian process can be obtained from our Bayesian linear regression model$f(x) = \phi(x)^T\omega$ with prior $\omega \sim \mathcal{N}(0, \Sigma_p)$. We have for the mean and covariance

$$\begin{aligned} \mathrm{E}[f(x)] &= \phi(x)^T \mathrm{E}[\omega] = 0 \\ \mathrm{E}[f(x)f(x)'] &= \phi(x)^T \mathrm{E}[\omega\omega^T]\phi(x') = \phi(x)^T \Sigma_p \phi(x') \end{aligned}$$

- The marginalization property simply means that if the GP e.g. specifies $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$, then it must also specify $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ where Σ_{11} is the relevant submatrix of Σ

- A simple example of a Gaussian process can be obtained from our Bayesian linear regression model $f(x) = \phi(x)^T\omega$ with prior $\omega \sim \mathcal{N}(0, \Sigma_p)$. We have for the mean and covariance

$$
\begin{aligned}
\mathrm{E}[f(x)] &= \phi(x)^T \mathrm{E}[\omega] = 0 \\
\mathrm{E}[f(x)f(x)'] &= \phi(x)^T \mathrm{E}[\omega\omega^T]\phi(x') = \phi(x)^T \Sigma_p \phi(x')
\end{aligned}
$$

Function-Space View

- In this module we consider the squared exponential (SE) covariance function. A list of covariance function is presented later

- The covariance function specifies the covariance between pairs of random variables

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\{-\frac{1}{2}|x_p - x_q|^2\}.$$

- Note, that the covariance between the outputs is written as a function of the inputs.

- In this module we consider the squared exponential (SE) covariance function. A list of covariance function is presented later

- The covariance function specifies the covariance between pairs of random variables

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\{-\frac{1}{2}|x_p - x_q|^2\}.$$

- Note, that the covariance between the outputs is written as a function of the inputs.

- In this module we consider the squared exponential (SE) covariance function. A list of covariance function is presented later

- The covariance function specifies the covariance between pairs of random variables

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\{-\frac{1}{2}|x_p - x_q|^2\}.$$

- Note, that the covariance between the outputs is written as a function of the inputs.

- For squared error covariance function, we see that the covariance is almost unity between variables whose corresponding inputs are very close, and decreases as their distance in the input space increases.

- It can be shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.

- For every positive definite covariance function $k(,)$, there exists a (possibly infinite) expansion in terms of basis functions, known as Mercer's theorem.

- We generate a random Gaussian vector with this covariance matrix

$$f_* \sim \mathcal{N}(0, K(X_*, X_*))$$

Function-Space View

- For squared error covariance function, we see that the covariance is almost unity between variables whose corresponding inputs are very close, and decreases as their distance in the input space increases.

- It can be shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.

- For every positive definite covariance function $k(,)$, there exists a (possibly infinite) expansion in terms of basis functions, known as Mercer's theorem.

- We generate a random Gaussian vector with this covariance matrix

$$f_* \sim \mathcal{N}(0, K(X_*, X_*))$$

Pathshala पाठशाला MHRD Govt. of India

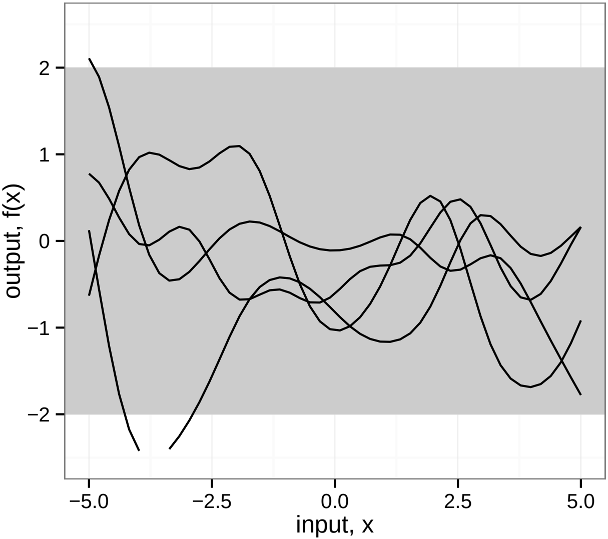
- For squared error covariance function, we see that the covariance is almost unity between variables whose corresponding inputs are very close, and decreases as their distance in the input space increases.
- It can be shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.
- For every positive definite covariance function $k(,)$, there exists a (possibly infinite) expansion in terms of basis functions, known as Mercer's theorem.
- We generate a random Gaussian vector with this covariance matrix

$$f_* \sim \mathcal{N}(0, K(X_*, X_*))$$

- For squared error covariance function, we see that the covariance is almost unity between variables whose corresponding inputs are very close, and decreases as their distance in the input space increases.
- It can be shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.
- For every positive definite covariance function $k(,)$, there exists a (possibly infinite) expansion in terms of basis functions, known as Mercer's theorem.
- We generate a random Gaussian vector with this covariance matrix

$$f_* \sim \mathcal{N}(0, K(X_*, X_*))$$

Function-Space View



- In the graph above - four functions drawn at random from a GP prior
- Notice that "informally" the functions look smooth.
- squared exponential covariance function is infinitely differentiable
- The characteristic length-scale is around one unit.

- In the graph above - four functions drawn at random from a GP prior
- Notice that “informally” the functions look smooth.
- squared exponential covariance function is infinitely differentiable
- The characteristic length-scale is around one unit.

- In the graph above - four functions drawn at random from a GP prior
- Notice that "informally" the functions look smooth.
- squared exponential covariance function is infinitely differentiable
- The characteristic length-scale is around one unit.

- In the graph above - four functions drawn at random from a GP prior
- Notice that “informally” the functions look smooth.
- squared exponential covariance function is infinitely differentiable
- The characteristic length-scale is around one unit.

- more realistic modelling situations that we do not have access to function values themselves, but only noisy versions

$$y = f(x) + \epsilon$$

- Assuming additive independent identically distributed Gaussian noise ϵ with variance σ_n^2 the prior on the noisy observations becomes

$$cov(y_p, y_q) = k(x_p, x_q) + \sigma_n^2 \delta_{pq}$$

or

$$cov(y) = K(X, X) + \sigma_n^2 \mathbf{I}$$

where δ_{pq} is kronecker delta which is one iff $p = q$ and zero otherwise.

- more realistic modelling situations that we do not have access to function values themselves, but only noisy versions

$$y = f(x) + \epsilon$$

- Assuming additive independent identically distributed Gaussian noise ϵ with variance σ_n^2 the prior on the noisy observations becomes

$$cov(y_p, y_q) = k(x_p, x_q) + \sigma_n^2 \delta_{pq}$$

or

$$cov(y) = K(X, X) + \sigma_n^2 \mathbf{I}$$

where δ_{pq} is kronecker delta which is one iff $p = q$ and zero otherwise.

Function-Space View

- Joint distribution of the observed target values and the function values at the test locations under the prior as

$$\begin{bmatrix} f_* \\ y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbf{I} & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right).$$

- Predictive equations for Gaussian process regression

$$f_* | X, y, X_* \sim \mathcal{N}\left(f_*, cov(f_*)\right)$$

where
$\mathrm{f}_* = \mathbb{E}[f_* | X, y, X_*] = K(X_*, X)[K(X,X) +_n^2 \mathbf{I}]^{-1} y$
$\mathrm{cov}(\mathrm{f}_*) =$
$K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma_n^2 \mathbf{I}]^{-1} K(X, X_*)$

Function-Space View

- Joint distribution of the observed target values and the function values at the test locations under the prior as

$$\begin{bmatrix} f_* \\ y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbf{I} & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right).$$

- Predictive equations for Gaussian process regression

$$f_*|X, y, X_* \sim \mathcal{N}\left(f_*, cov(f_*)\right)$$

where
$\mathsf{f}_* = \mathbb{E}[f_*|X, y, X_*] = K(X_*, X)[K(X,X) +_n^2 \mathbf{I}]^{-1}y$
$\mathsf{cov}(\mathsf{f}_*) =$
$K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma_n^2\mathbf{I}]^{-1}K(X, X_*)$

- The covariance functions that we use will have some free parameters.
- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.
- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.
- In general we call the free parameters hyperparameters.

Varying the Hyperparameters

- The covariance functions that we use will have some free parameters.
- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.
- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.
- In general we call the free parameters hyperparameters.

- The covariance functions that we use will have some free parameters.

- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.

- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.

- In general we call the free parameters hyperparameters.

Varying the Hyperparameters

- The covariance functions that we use will have some free parameters.

- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.

- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.

- In general we call the free parameters hyperparameters.

Varying the Hyperparameters

- The covariance functions that we use will have some free parameters.
- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.
- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.
- In general we call the free parameters hyperparameters.

Varying the Hyperparameters

- The covariance functions that we use will have some free parameters.
- For example, the squared-exponential covariance function in one dimension has the following form

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

- The covariance is denoted k_y as it is for the noisy targets y rather than for the underlying function f.
- Observe that the length-scale l, the signal variance σ_f^2 and noise variance σ_n^2 can be varied.
- In general we call the free parameters hyperparameters.

- Estimation of these hyperparameters are an important steps.
- The marginal likelihood is the integral of the likelihood times the prior
$$p(y|X) = \int p(y|f, X)p(f|X)df.$$
- The term marginal likelihood refers to the marginalization over the function values f.
- Under the Gaussian process model the prior is Gaussian, $y|f \sim \mathcal{N}(f, \sigma_n^2 \mathbf{I})$ or
$$\log p(y|X, \theta, \sigma_n^2) = -\frac{1}{2} y^T (K_\theta + \sigma_n^2 \mathbf{I})^{-1} y - \frac{1}{2} \log |K_\theta + \sigma_n^2 \mathbf{I}| - \frac{n}{2} \log 2\pi$$
- This result can also be obtained directly by observing that $y \sim \mathcal{N}(0, K_\theta + \sigma_n^2 \mathbf{I})$

- Estimation of these hyperparameters are an important steps.
- The marginal likelihood is the integral of the likelihood times the prior
$$p(y|X) = \int p(y|f, X)p(f|X)df.$$
- The term marginal likelihood refers to the marginalization over the function values f.
- Under the Gaussian process model the prior is Gaussian, $y|f \sim \mathcal{N}(f, \sigma_n^2 \mathbf{I})$ or
$$\log p(y|X, \theta, \sigma_n^2) = -\frac{1}{2} y^T (K_\theta + \sigma_n^2 \mathbf{I})^{-1} y - \frac{1}{2} \log |K_\theta + \sigma_n^2 \mathbf{I}| - \frac{n}{2} \log 2\pi$$
- This result can also be obtained directly by observing that $y \sim \mathcal{N}(0, K_\theta + \sigma_n^2 \mathbf{I})$

- Estimation of these hyperparameters are an important steps.

- The marginal likelihood is the integral of the likelihood times the prior
$$p(y|X) = \int p(y|f, X)p(f|X)df.$$

- The term marginal likelihood refers to the marginalization over the function values f.

- Under the Gaussian process model the prior is Gaussian, $y|f \sim \mathcal{N}(f, \sigma_n^2\mathbf{I})$ or
$$\log p(y|X, \theta, \sigma_n^2) = -\frac{1}{2}y^T(K_\theta+\sigma_n^2\mathbf{I})^{-1}y - \frac{1}{2}\log|K_\theta+\sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi$$

- This result can also be obtained directly by observing that $y \sim \mathcal{N}(0, K_\theta + \sigma_n^2\mathbf{I})$

Estimation of Hyperparameters

- Estimation of these hyperparameters are an important steps.

- The marginal likelihood is the integral of the likelihood times the prior

$$p(y|X) = \int p(y|f, X)p(f|X)df.$$

- The term marginal likelihood refers to the marginalization over the function values f.

- Under the Gaussian process model the prior is Gaussian, $y|f \sim \mathcal{N}(f, \sigma_n^2\mathbf{I})$ or

$$\log p(y|X, \theta, \sigma_n^2) = -\frac{1}{2}y^T(K_\theta + \sigma_n^2\mathbf{I})^{-1}y - \frac{1}{2}\log|K_\theta + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi$$

- This result can also be obtained directly by observing that $y \sim \mathcal{N}(0, K_\theta + \sigma_n^2\mathbf{I})$

log-Posterior for Hyperparameters

- We can assume suitable hyper-prior on (θ, σ_n^2) as $(\theta, \sigma_n^2) = p(\theta)p(\sigma_n^2)$

- The log-posterior distribution is

$$\log p(\theta, \sigma_n^2 | y, X) \propto \log p(y|X, \theta, \sigma_n^2) + \log p(\theta) + \log p(\sigma_n^2)$$

- The posterior mode is

$$\hat{\theta} = \underset{\theta \in \Theta}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

and

$$\hat{\sigma_n^2} = \underset{\sigma_n^2 \in \mathbb{R}^+}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

log-Posterior for Hyperparameters

- We can assume suitable hyper-prior on (θ, σ_n^2) as $(\theta, \sigma_n^2) = p(\theta)p(\sigma_n^2)$

- The log-posterior distribution is

$$\log p(\theta, \sigma_n^2 | y, X) \propto \log p(y|X, \theta, \sigma_n^2) + \log p(\theta) + \log p(\sigma_n^2)$$

- The posterior mode is

$$\hat{\theta} = \underset{\theta \in \Theta}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

and

$$\hat{\sigma_n^2} = \underset{\sigma_n^2 \in \mathbb{R}^+}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

- We can assume suitable hyper-prior on (θ, σ_n^2) as $(\theta, \sigma_n^2) = p(\theta)p(\sigma_n^2)$

- The log-posterior distribution is

$$\log p(\theta, \sigma_n^2 | y, X) \propto \log p(y|X, \theta, \sigma_n^2) + \log p(\theta) + \log p(\sigma_n^2)$$

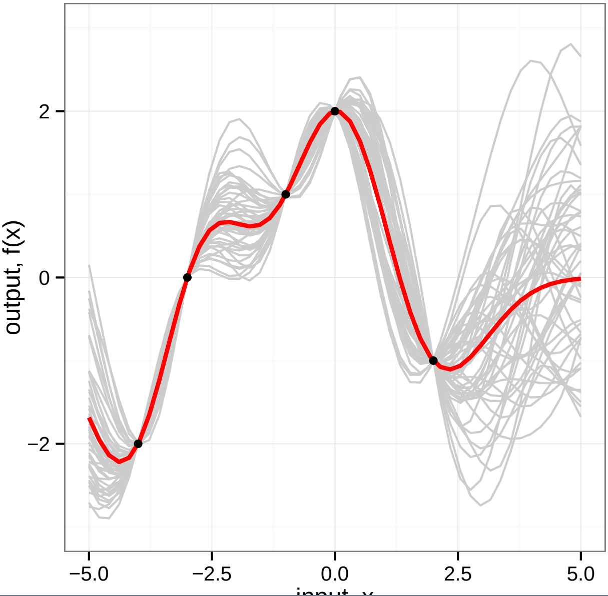
- The posterior mode is

$$\hat{\theta} = \underset{\theta \in \Theta}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

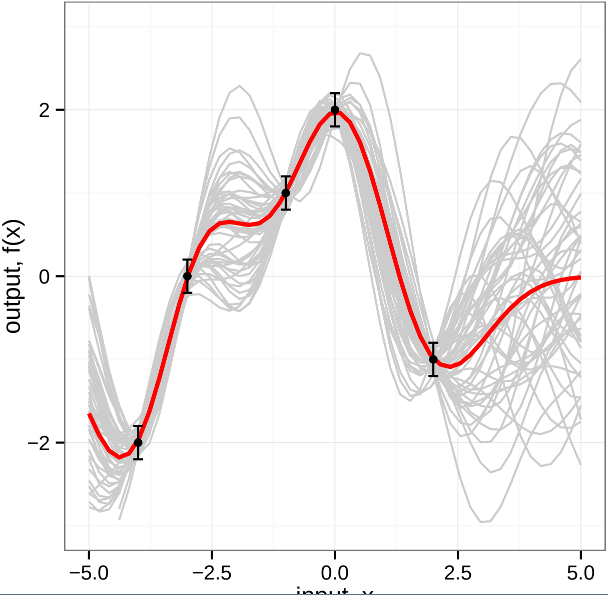
and

$$\hat{\sigma_n^2} = \underset{\sigma_n^2 \in \mathbb{R}^+}{argmax} \quad \log p(\theta, \sigma_n^2 | y, X)$$

Estimated Curve



Estimated Curve with CI



Maximum Likelihood Estimation - The Basic Idea

Module 1

Saurav De

Department of Statistics
Presidency University

Maximum Likelihood Estimation - The Basic Idea

- Let us begin with the following example ...
- **Ex.1** A certain lion has three possible state of activities, namely Very active, Moderately active and Lethargic.
  So state of activity is the **parameter**, say θ which has three known levels, say $\theta_1 =$ Very active, $\theta_2 =$ Moderately active, and $\theta_3 =$ Lethargic.
- So we can say that the (known) parameter space $\Omega = \{\theta_1, \theta_2, \theta_3\}$.

Maximum Likelihood Estimation - The Basic Idea

- Suppose, each night this lion eats X animals with probability $P_\theta[X = x] = f(x, \theta)$, say. Then consider the following table.

x	0	1	2	3	4
$P_{\theta_1}[X = x]$	0	0.05	0.05	0.8	0.1
$P_{\theta_2}[X = x]$	0.05	0.05	0.8	0.1	0
$P_{\theta_3}[X = x]$	0.9	0.09	0.01	0	0

Maximum Likelihood Estimation - The Basic Idea

- From the above table we can say, if x is kept fixed at the value 4, then the most likely value of θ is θ_1 (*last column of the table*).
- If x is given as 1, then θ_3 is the most likely value of θ (*second column of the table*), etc.

- The estimate of θ, so obtained, is well-known as the **Maximum Likelihood (ML) Estimate** of θ.
- The method, applied to get such estimate, is known as the **Maximum Likelihood Method of Estimation**.

Maximum Likelihood Estimation - The Basic Idea

Let us now formalise this celebrated method of estimation through the following discussion.

Maximum Likelihood Estimation - The Basic Idea

- Suppose $\{p(\mathbf{x}, \boldsymbol{\theta}); \boldsymbol{\theta} \in \Omega\}$ = A family of joint probability function (p.m.f. or p.d.f.) of $(X_1, \ldots, X_n)$ characterised by the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)'; s \geq 1$, and Ω = the parameter space.
- Given $\mathbf{x}$, the corresponding family member becomes a function of $\boldsymbol{\theta}$, called the **likelihood function** of $\boldsymbol{\theta}$ given $\mathbf{x}$ and is usually denoted by $L_{\mathbf{x}}(\boldsymbol{\theta})$ or simply $L(\boldsymbol{\theta}); \boldsymbol{\theta} \in \Omega$.
- So in the above table rows give the p.m.f. values of the distribution of X under different fixed choices of θ; whereas the first column gives the values of $L(\theta)$ for $\theta = \theta_1, \theta_2, \theta_3$ when $x = 0$, column 2 provides the same given $x = 1$ etc.

Maximum Likelihood Estimation - The Basic Idea

Now any value of $\boldsymbol{\theta}$ (*within parameter space*) for which the likelihood fuction attains its supremum, is known as the Maximum Likelihood Estimate of $\boldsymbol{\theta}$. So in the above example θ_1 is the ML estimate of θ given $x = 4$, θ_2 is the ML estimate of θ given $x = 2$ etc.

Maximum Likelihood Estimation - The Basic Idea

Thus the principle of maximum likelihood estimation consists of selecting as an estimator of θ (*real or vector-valued*) a $\hat{\theta}(\mathbf{X})$ (where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$) that maximises the likelihood function $L(\theta)$, or in other word, to find a mapping $\hat{\theta}$ of $\mathcal{R}_n \longrightarrow \Omega$ that satisfies

$$L(\hat{\theta}) = \sup_{\theta \in \Omega} L(\theta).$$

Such a $\hat{\theta}$, if exists, is called Maximum Likelihood Estimator.

Maximum Likelihood Estimation - The Basic Idea

Some Introductory Examples

Example A. Let $X_1, X_2, \ldots, X_n$ be *iid* $R(0, \theta)$ variables; $\theta \in (0, \infty)$.

Given $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the likelihood function

$$
\begin{aligned}
L_{\mathbf{x}}(\theta) &= \frac{1}{\theta^n}, 0 \leq x_i \leq \theta; i = 1, 2, \ldots, n \Leftrightarrow \theta \geq x_{(n)} \\
&= 0, \quad \text{otherwise}
\end{aligned}
$$

where $x_{(n)} = max \ \{x_1, x_2, \ldots, x_n\}$.

So here $\hat{\theta}_{MLE} = X_{(n)} \in (0, \infty)$

Maximum Likelihood Estimation - The Basic Idea

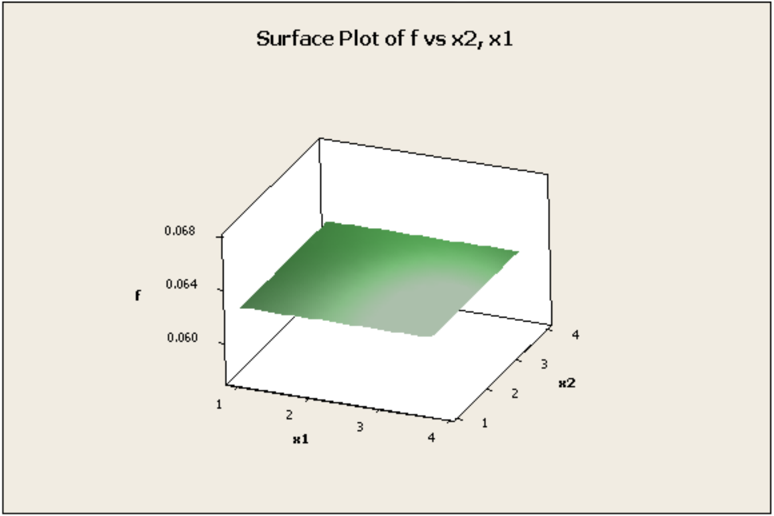
A note on Ex A.

Let $n = 2$ i.e. $X_1, X_2 \sim R(0, \theta)$ independently.

- The joint pdf of X_1, X_2 is a bivariate function

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\theta^2} \ \ 0 \leq x_1, x_2 \leq \theta \\ &= 0 \ \ \text{o.w.} \end{aligned}$$

Maximum Likelihood Estimation - The Basic Idea



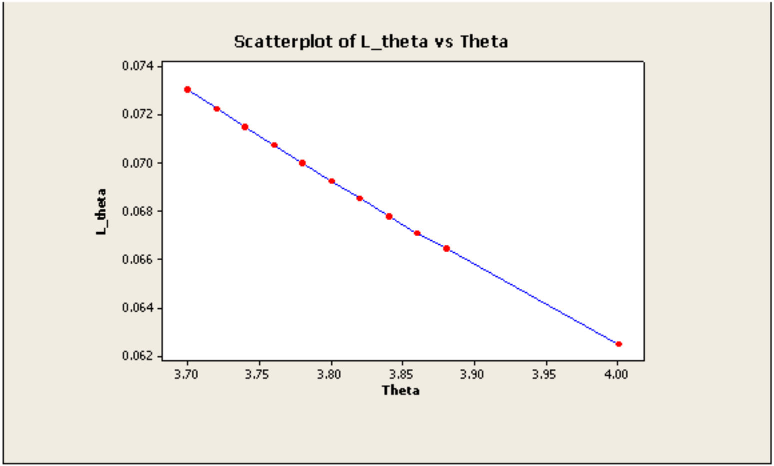
Maximum Likelihood Estimation - The Basic Idea

But the likelihood function for given x_1, x_2 is a univariate function

$$
\begin{aligned}
L(\theta) &= \frac{1}{\theta^2} \;\; x_{(2)} \leq \theta < \infty \\
&= 0 \;\; \text{o.w.}
\end{aligned}
$$

The corresponding plot follows in the next slide.

Maximum Likelihood Estimation - The Basic Idea



Maximum Likelihood Estimation - The Basic Idea

A similar type of illustration under normal family of distributions using R.

R Code :

```
> library(TeachingDemos)
> m <- runif(1, 50,100)#initial value of mean
> s <- runif(1, 1, 10)#initial value of SD
> x <- rnorm(15, m, s)#initial sample drawn from normal
    distribution
> mm <- mean(x)
> ss <- sqrt(var(x))
> ss2 <- sqrt(var(x)*11/12)
> mle.demo(x)#log likelihood calculated here
```

Maximum Likelihood Estimation - The Basic Idea

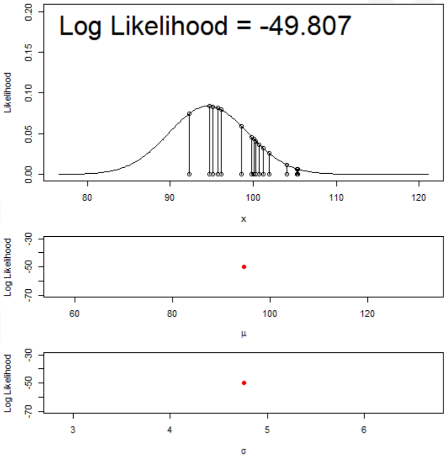


Figure: Log Likelihood when $\mu = 94.6$, $\sigma = 4.8$

Maximum Likelihood Estimation - The Basic Idea

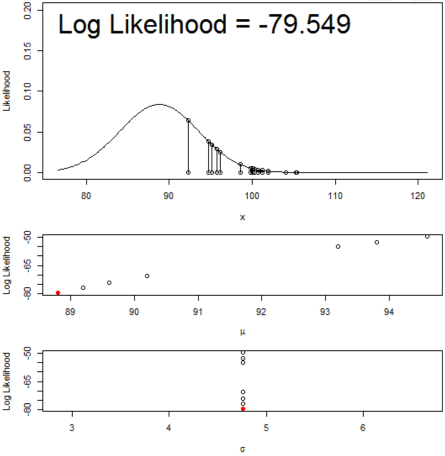


Figure: Log Likelihood when $\mu = 88.8$, $\sigma = 4.8$

Maximum Likelihood Estimation - The Basic Idea

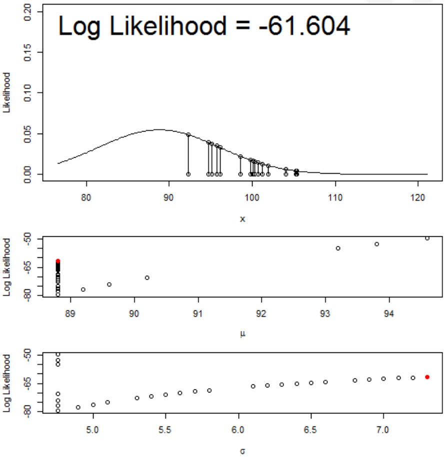


Figure: Log Likelihood when $\mu = 88.8$, $\sigma = 7.3$

Maximum Likelihood Estimation - The Basic Idea

Ex B. Let $X_1, X_2, \ldots, X_n$ be *iid* with common pdf

$$\begin{aligned} f_\theta(x) &= \theta x^{\theta-1} \ ; \ 0 < x < 1 \\ &= 0 \ \text{otherwise} \ ; \ \theta > 0 \end{aligned}$$

Then, given $\mathbf{x}$, the likelihood function is

$$L(\theta) = \theta^n (\prod_{i=1}^{n} x_i)^{\theta-1}$$

$$\Rightarrow \ell(\theta) = \log L(\theta) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i$$

$$\frac{\partial}{\partial\theta} \ell(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log x_i$$

Maximum Likelihood Estimation - The Basic Idea

By maxima-minima principle $\frac{\partial}{\partial\theta}\ell(\theta) = 0 \Rightarrow \hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log x_i} \in (0, \infty)$

Also S.O.C. $\frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\frac{n}{\theta^2} < 0 \ \forall \ \theta > 0$ ensures that $\hat{\theta}$ maximises the likelihood function and hence the MLE of θ.

Maximum Likelihood Estimation - The Basic Idea

Ex C. Let $X_1, X_2, \ldots, X_n$ be *iid* Bernoulli (θ) random variables, $\theta \in [0,1]$. Then, given $\mathbf{x}$, the likelihood function is

$$L(\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}$$

Let $\sum_{i=1}^{n} x_i = a$.

Suppose

Case I: $\underline{0 < a < n}$ i.e. $\underline{0 < \frac{a}{n} < 1}$

Case II: $\underline{a = 0}$ i.e. $\underline{\frac{a}{n} = 0}$

Case III: $\underline{a = n}$ i.e. $\underline{\frac{a}{n} = 1}$

Maximum Likelihood Estimation - The Basic Idea

Case I: $\underline{0 < a < n}$

For the n numbers $\frac{n\theta}{a}, \ldots, \frac{n\theta}{a}$ (a times) and $\frac{n(1-\theta)}{n-a}, \ldots, \frac{n(1-\theta)}{n-a}$ ($n-a$ times) AM-GM inequality gives

$$\frac{a\frac{n\theta}{a} + (n-a)\frac{n(1-\theta)}{n-a}}{n} \geq \left[\left(\frac{n\theta}{a}\right)^a \left(\frac{n(1-\theta)}{n-a}\right)^{n-a}\right]^{1/n}$$

$$\Leftrightarrow \theta^a(1-\theta)^{n-a} \leq \left(\frac{a}{n}\right)^a \left(1-\frac{a}{n}\right)^{n-a}$$

Or

$$L(\theta) \leq \left(\frac{a}{n}\right)^a \left(1-\frac{a}{n}\right)^{n-a} \; ; \; 0 < \frac{a}{n} < 1$$

Maximum Likelihood Estimation - The Basic Idea

Case II: $\underline{a = 0}$

$a = 0 \Leftrightarrow x_i = 0, i = 1, 2, \ldots, n$ which means $L(\theta) = (1 - \theta)^n$ which is supremum at $\theta = 0 = \frac{a}{n}$.

Case III: $\underline{a = n}$

$a = n \Leftrightarrow x_i = 1, i = 1, 2, \ldots, n$ which means $L(\theta) = \theta^n$ which is supremum at $\theta = 1 = \frac{a}{n}$.
So under all the cases, $L(\theta)$ atains its supremum at $\theta = \frac{a}{n} \in [0, 1]$.
$\Rightarrow$ the MLE of θ is $\hat{\theta} = \frac{a}{n} = \overline{x}$; the sample mean.

Maximum Likelihood Estimation - The Basic Idea

Application 1. Let $X \sim$ Hypergeometric distribution and given $X = x$, the likelihood function be

$$L(N) = \frac{{}^{M}C_{x} \; {}^{N-M}C_{n-x}}{{}^{N}C_{n}} \quad N = 2, 3, \ldots$$

a function of the parameter N; M and n being known.

Let $R(N) = \frac{L(N)}{L(N-1)} = \frac{(N-n)(N-M)}{N(N-M-n+x)}$.

Now, for values of N for which the ratio $R(N) > 1$, $L(N)$ increases with N and $L(N)$ is decresing in N for those N which results in $R(N) < 1$.

Maximum Likelihood Estimation - The Basic Idea

Let $\hat{N} = \hat{N}(X)$ be the MLE of N. Then

$R(\hat{N}) \geq (\leq) 1$ if and only if $\hat{N} \leq (\geq) \frac{nM}{x}$

This implies that $L(N)$ attains its maximum at $N \approx \frac{nM}{x}$. Thus $\hat{N}(X) = \left[\frac{nM}{X}\right]$ where $[x]$ denotes the largest integer contained in x.

Maximum Likelihood Estimation - The Basic Idea

Application 2. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ common pdf $f_\theta(x)$ where

$$\begin{aligned} f_\theta(x) &= \frac{2x}{2\alpha}, \qquad 0 \leq x \leq \theta \\ &= \frac{2(\alpha - x)}{\alpha(\alpha - \theta)}, \quad \theta \leq \alpha \\ &= 0, \text{otherwise}, \end{aligned}$$

where $\alpha > 0$ is a (known) constant and $\theta \in (0, \alpha)$. Here the likelihood function of θ is

$$L(\theta) = \left(\frac{2}{\alpha}\right)^n \prod_{x_i \leq \theta} \frac{x_i}{\theta} \prod_{x_i > \theta} \frac{\alpha - x_i}{\alpha - \theta}$$

Assume that $0 \leq x_{(1)} < x_{(2)} < \ldots < x_{(n)} \leq \alpha$, i.e. the ordered observations .

Maximum Likelihood Estimation - The Basic Idea

case I: $x_{(j)} < \theta < x_{(j+1)}$

Then $L(\theta) = \left(\frac{2}{\alpha}\right)^n \theta^{-j}(\alpha - \theta)^{-(n-j)} \prod_{i=1}^{j} x_{(i)} \prod_{i=j+1}^{n} (\alpha - x_{(i)})$

Check that the **first-order** and **second-order** condition ensure that the stationary value of θ that exists must be a point of minimum, $\forall\ j = 1, 2, \ldots, n-1$.

Maximum Likelihood Estimation - The Basic Idea

case II: $0 \leq \theta \leq x_{(1)}$

Then $L(\theta) = \left(\frac{2}{\alpha}\right)^n (\alpha - \theta)^{-n} \prod_{i=1}^{n} (\alpha - x_{(i)}) \uparrow \theta.$

Similarly under **case III:** $x_{(n)} \leq \theta \leq \alpha$, $L(\theta) \downarrow \theta$.

So the ML estimate of θ is either x_1 or x_n depending on the observations and value of α.

Note. This is an example where MLE is necessarily an actual observation, but not necessarily any particular observation.

Maximum Likelihood Estimation - The Basic Idea

Application 3. Let $X_1, X_2, \ldots, X_n$ be a random sample from some DF F on the real line. Suppose that we observe $x_1, x_2, \ldots, x_n$ which are all different. show that the MLE of F is F_n, the empirical DF of the sample.

Proof. Empirical DF at the point x is defined as $F_n(x) = \frac{\#\{X_i \leq x\}}{n}$. Let $F(x) = \theta$, the parameter of the distribution.

Define $Y_i = 1(0)$ for $X_i \leq x$ $(X_i > x)$, $i = 1, 2, \ldots, n$. Then $P_\theta[Y_i = 1] = P_\theta[X_i \leq x] = F(x) = \theta$. So $Y_1, Y_2, \ldots, Y_n$ follow Bernoulli(θ), $0 \leq \theta \leq 1$; independently.

Then the MLE of θ is $\overline{Y} = \sum_{i=1}^{n} Y_i/n = \frac{\#\{X_i \leq x\}}{n}$. *Proved*

TRY YOURSELF!

M1.1. A random sample of size n be drawn from a population with density

$$\begin{aligned} f_\theta(x) &= \frac{\theta}{x^{\theta+1}} \quad 1 \le x < \infty \\ &= 0 \quad \text{o.w.} \end{aligned}$$

Find MLE of θ if exists.

M1.2. Find MLE of θ based on a random sample of size n for the following Rectangular distributions

a. $R(-\theta, \theta)$ $\theta > 0$

b. $R(\theta, \theta^2)$ $\theta > 1$

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 1 . . .

M1.1. The positive joint density of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$f_\theta(\mathbf{x}) = \frac{\theta^n}{\prod_{i=1}^{n} x_i^{\theta+1}} \; ; x_i \geq 1 \; \forall \; i$$

So the likelihood function of θ for given $\mathbf{x}$ is

$$L(\theta) = \frac{\theta^n}{\prod_{i=1}^{n} x_i^{\theta+1}} \; , \; 0 < \theta < \infty$$

Maximum Likelihood Estimation - The Basic Idea

Here $L(\theta)$ may not be a monotone function of θ.
Using maxima-minima principle we can find that $L(\theta)$ gets maximised for the value of θ as $\frac{n}{\sum_{i=1}^{n} \log x_i} \in (0, \infty)$.

Hence the MLE of θ is $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i}$

Maximum Likelihood Estimation - The Basic Idea

M1.2. a. The joint pdf of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from $R(-\theta, \theta)$ $\theta > 0$ is

$$\begin{aligned} f_\theta(\mathbf{x}) &= \frac{1}{(2\theta)^n} \quad -\theta \leq x_1, x_2, \ldots, x_n \leq \theta \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Or

$$\begin{aligned} f_\theta(\mathbf{x}) &= \frac{1}{(2\theta)^n} \quad -\theta \leq x_{(1)} \; < \; x_{(n)} \leq \theta \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Now $-\theta \leq x_{(1)} \; , \; x_{(n)} \leq \theta \iff \theta \geq \max\left\{-x_{(1)} \; , \; x_{(n)}\right\}$.

Maximum Likelihood Estimation - The Basic Idea

So the likelihood function of θ is

$$\begin{aligned} L(\theta) &= \frac{1}{(2\theta)^n} \quad \max\left\{-x_{(1)}\ ,\ x_{(n)}\right\} \leq \theta < \infty \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Hence $L(\theta) \downarrow \theta$ for $0 < \max\left\{-x_{(1)}\ ,\ x_{(n)}\right\} \leq \theta < \infty$.

$\Longrightarrow$ MLE of θ is $\hat{\theta} = \max\left\{-X_{(1)}\ ,\ X_{(n)}\right\}$

Maximum Likelihood Estimation - The Basic Idea

M1.2. b. The likelihood function of θ based on $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from $R(\theta, \theta^2)$ $\theta > 1$ is

$$\begin{aligned} L(\theta) &= \frac{1}{\theta(\theta-1)} \quad \theta \leq x_{(1)} < x_{(n)} \leq \theta^2 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

That is

$$\begin{aligned} L(\theta) &= \frac{1}{\theta(\theta-1)} \quad 1 < \sqrt{x_{(n)}} \leq \theta \leq x_{(1)} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Now it is easy to check that under $1 < \sqrt{x_{(n)}} \leq \theta \leq x_{(1)}$, $L(\theta) > 0$ and $\downarrow \theta$. So the supremum of $L(\theta)$ attains for $\hat{\theta} = \sqrt{x_{(n)}}$. $\implies$ the MLE of θ is $\sqrt{X_{(n)}}$

MLE on Truncated Parameter Space

Module 2

Saurav De

Department of Statistics
Presidency University

MLE on Truncated Parameter Space

Let us begin with an illustration.

Suppose from a Poisson (λ) distribution 10 randomly chosen sample observations are as follows

2 3 4 1 2 1 2 1 3 0

If the sample observations are denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_{10})$, then given $\mathbf{x}$ the likelihood function of λ is

$$L(\lambda) = \frac{\exp[-10\lambda]\lambda^{10\overline{x}}}{\prod_{i=1}^{10} x_i!} = \frac{\exp[-10\lambda]\lambda^{19}}{6912}$$

where $\overline{x}$ = sample mean = 1.9 and $\prod_{i=1}^{10} x_i! = 6912$ from the fiven data.

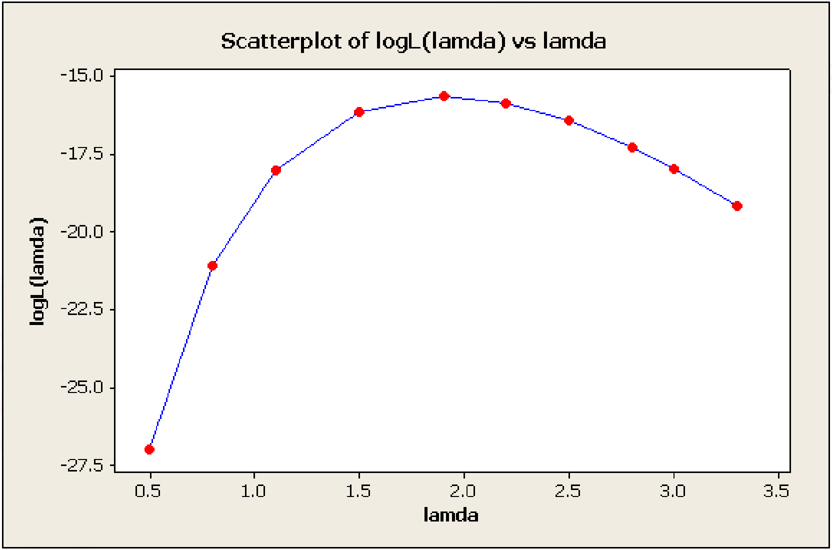
MLE on Truncated Parameter Space

A simple plot of $\log L(\lambda)$ against nonnegative values of λ based on the following table

λ	$\log L(\lambda)$
0.5	-27.0108
0.8	-21.0807
1.1	-18.0301
1.5	-16.1372
1.9	-15.6458
2.2	-15.8603
2.5	-16.4315
2.8	-17.2782
3.0	-17.9674
3.3	-19.1565

is as follows:

MLE on Truncated Parameter Space



We know for nonnegative λ of Poisson probability model the ML estimate is $\overline{x}$ (*try yourself*), i.e. here 1.9. This is evident also from the plot that the log of likelihood attains its maximum at 1.9 (very close to 2).

Now, suppose instead of $[0, \infty)$ if the parameter space is like $[\lambda_0, \infty)$, λ_0, known, the following cases are obvious to realise at least from the plot.

<u>case I:</u> If $\lambda_0 \leq \overline{x} < \infty$, $L(\lambda)$ is still maximum at $\overline{x} \implies$ MLE $(\hat{\lambda}) = \overline{x}$

MLE on Truncated Parameter Space

<u>case II:</u> If $\lambda_0 > \overline{x}$, $L(\lambda) \downarrow \lambda \implies \hat{\lambda} = \lambda_0$

$$\begin{aligned} \hat{\lambda} &= \lambda_0 \quad \text{if} \quad \overline{x} < \lambda_0 \\ &= \overline{x} \quad \text{if} \quad \overline{x} \in [\lambda_0, \infty) \end{aligned}$$

Similarly, if the parameter space is like $[\lambda_0, \lambda_1]$, λ_0 , λ_1, known, the following cases are obvious to realise at least from the plot.

<u>case I:</u> If $\lambda_0 > \overline{x}$, $L(\lambda) \downarrow \lambda \implies \hat{\lambda} = \lambda_0$

<u>case II:</u> If $\lambda_1 < \overline{x}$, $L(\lambda) \uparrow \lambda \implies \hat{\lambda} = \lambda_1$

MLE on Truncated Parameter Space

case III: If $\lambda_0 \leq \overline{x} \leq \lambda_1$, $L(\lambda)$ is still maximum at $\overline{x} \implies$ MLE $(\hat{\lambda}) = \overline{x}$

$$\begin{aligned} \hat{\lambda} &= \lambda_0 \quad \text{if} \quad \overline{x} < \lambda_0 \\ &= \overline{x} \quad \text{if} \quad \overline{x} \in [\lambda_0, \lambda_1] \\ &= \lambda_1 \quad \text{if} \quad \overline{x} > \lambda_1 \end{aligned}$$

MLE on Truncated Parameter Space

Next we consider another illustration with Exponential (mean $= \theta$) distribution having density

$$
\begin{aligned}
f_\theta(x) &= \frac{1}{\theta}\exp(-\frac{x}{\theta}) \ , \quad x > 0 \\
&= 0 \ \text{ otherwise}
\end{aligned}
$$

where $\theta \in (0, \infty)$.

Based on a random sample $X_1, X_2, \ldots, X_n$ the likelihood is

$$L(\theta) = \tfrac{1}{\theta^n}\exp(-\sum_{i=1}^{n} x_i/\theta) \ , \ \theta \in (0, \infty)$$

Using maxima-minima principle, $L(\theta)$ attains its maximum at sample mean $\overline{x}$.

$\Longrightarrow$ MLE $(\hat{\theta}) = \overline{X}$.

Now, suppose instead of $(0, \infty)$ if the parameter space is like $[\theta_0, \infty)$, θ_0, known, the following cases are obvious to realise at least from the plot.

case I: If $\theta_0 \leq \overline{x} < \infty$, $L(\theta)$ is still maximum at $\overline{x} \implies$ MLE $(\hat{\theta}) = \overline{x}$

case II: If $\theta_0 > \overline{x}$, $L(\theta) \downarrow \theta \implies \hat{\theta} = \theta_0$

$$\begin{aligned} \hat{\theta} &= \theta_0 \quad \text{if} \quad \overline{x} < \theta_0 \\ &= \overline{x} \quad \text{if} \quad \overline{x} \in [\theta_0, \infty) \end{aligned}$$

MLE on Truncated Parameter Space

Similarly, if the parameter space is like $[\theta_0, \theta_1]$, θ_0 , θ_1, known, the following cases are obvious to realise at least from the plot.

<u>case I:</u> If $\theta_0 > \overline{x}$, $L(\theta) \downarrow \theta \implies \hat{\theta} = \theta_0$

<u>case II:</u> If $\theta_1 < \overline{x}$, $L(\theta) \uparrow \theta \implies \hat{\theta} = \theta_1$

<u>case III:</u> If $\theta_0 \leq \overline{x} \leq \theta_1$, $L(\theta)$ is still maximum at $\overline{x} \implies$ MLE $(\hat{\theta}) = \overline{x}$

$$\begin{aligned} \hat{\theta} &= \theta_0 \quad \text{if} \quad \overline{x} < \theta_0 \\ &= \overline{x} \quad \text{if} \quad \overline{x} \in [\theta_0, \theta_1] \\ &= \theta_1 \quad \text{if} \quad \overline{x} > \theta_1 \end{aligned}$$

MLE on Truncated Parameter Space

An illustration of the computation of MLE on truncated parameter space has been done on exponential distribution using R.

R Code and Output :

```
> theta0=2#lower bound of parametric space
> theta1=3#upper bound of parametric space
> theta=runif(1,theta0,theta1)
> theta#value of the parameter that is considered
[1] 2.896314
> samp=rexp(1000,(1/theta))#random sample drawn
> x_bar=mean(samp)
> x_bar#mean of the random sample drawn
[1] 2.848733
```

MLE on Truncated Parameter Space

R Code and Output (continued) :

```
> if(x_bar > theta1)
 {
 cat("\n")
 cat("MLE of 'theta' = ",theta1,"\n")
 }
> if(x_bar < theta0)
 {
 cat("\n")
 cat("MLE of 'theta' = ",theta0,"\n")
 }
 if(x_bar >= theta0 & x_bar <= theta1)
 {
 cat("\n")
 cat("MLE of 'theta' = ",x_bar,"\n")
 }
MLE of 'theta' =  2.848733
```

The discussion, that is done so far on the MLE under truncated parameter space, is a quite informal way. But based on this much, it is enough to state that in general if $T = T(\mathbf{X})$ be the MLE of a parameter θ under full or unrestricted parameter space, then under the truncated parameter space like $[\theta^*, \theta^{**}]$

$$
\begin{aligned}
\hat{\theta}(\mathbf{X}) &= \theta^* \text{ if } T(\mathbf{X}) < \theta^* \\
&= T(\mathbf{X}) \text{ if } \theta^* \leq T(\mathbf{X}) \leq \theta^{**} \\
&= \theta^{**} \text{ if } T(\mathbf{X}) > \theta^{**}
\end{aligned}
$$

Formal discussion on MLE based on truncated parameter space

Illustration. Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$ where $\mu \in [0, \infty)$. The the log of the likelihood of μ is

$$\log L(\mu) = \text{ constant } - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$\frac{\partial}{\partial\mu} \log L(\mu) = n(\overline{x} - \mu)$ and $\frac{\partial^2}{\partial\mu^2} \log L(\mu) = -n$ for any $\mu > 0$.

At $\mu = 0$ we consider differentiation with respect to μ from right hand side.

If $\overline{x} \in [0, \infty)$, from F.O.C. and S.O.C. we have MLE $\hat{\mu} = \overline{x}$

$$\overline{x} < 0 \implies \frac{\partial}{\partial\mu} \log L(\mu) < 0 \ \forall \ \mu \geq 0 \iff \log L(\mu) \downarrow \mu \ \forall \ \mu \geq 0.$$

$\implies$ $L(\mu)$ is maximised at $\mu = 0$. Therefore

$$\begin{aligned} \hat{\mu} &= \overline{x} \text{ if } \overline{x} \geq 0 \\ &= 0 \text{ if } \overline{x} < 0 \end{aligned}$$

- Asymptotic distribution of $\hat{\mu}$ when $\mu > 0$:

$$\sqrt{n}(\hat{\mu} - \mu) = \sqrt{n}(\overline{x} - \mu) + \sqrt{n}(\hat{\mu} - \overline{x}) \qquad \odot$$

Now $\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{D}} N(0,1)$ for any $\mu \in [0, \infty)$

And as $\hat{\mu} - \overline{X} = 0 \implies |\sqrt{n}(\hat{\mu} - \overline{X})| < \epsilon$ so

$P[|\sqrt{n}(\hat{\mu} - \overline{X})| < \epsilon] \geq P[\hat{\mu} - \overline{X} = 0] = P[\overline{X} \geq 0] = \Phi(\sqrt{n}\mu) \longrightarrow 1$ as $n \longrightarrow \infty$.

$$\implies \hat{\mu} - \overline{X} \xrightarrow{\mathcal{P}} 0$$

Finally $\odot \implies \sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{D}} N(0,1)$

MLE on Truncated Parameter Space

- Asymptotic distribution of $\hat{\mu}$ when $\mu = 0$:

Here $P[\overline{X} \geq 0] = \frac{1}{2} \; \forall \; n$

Thus the distribution function F_n of $\sqrt{n}(\hat{\mu} - 0)$ is defined as

$$\begin{aligned} F_n(y) &= 0 \text{ if } y < 0 \\ &= \Phi(\sqrt{n}y) \text{ if } y \geq 0 \end{aligned}$$

Note. This is not normal. Thus the asymptotic distribution of the MLE is normal for all $\mu \in [0, \infty)$ except at the boundary point $\mu = 0$ of the truncated parameter space.

Comparison through MSE criterion: MSE $(\hat{\mu}) \leq$ MSE $(\overline{X})$ for all $\mu \in [0, \infty)$

Justification. MSE $(\hat{\mu}) = \int_0^{\infty} (\overline{x} - \mu)^2 f_{\overline{X}}(\overline{x}) d\overline{x} + \int_{-\infty}^{0} \mu^2 f_{\overline{X}}(\overline{x}) d\overline{x}$

MSE $(\overline{X}) = \int_0^{\infty} (\overline{x} - \mu)^2 f_{\overline{X}}(\overline{x}) d\overline{x} + \int_{-\infty}^{0} (\overline{x} - \mu)^2 f_{\overline{X}}(\overline{x}) d\overline{x}$

$\implies$ MSE $(\overline{X})$ - MSE $(\hat{\mu}) = \int_{-\infty}^{0} \overline{x}(\overline{x} - 2\mu) f_{\overline{X}}(\overline{x}) d\overline{x}$

e-Pathshala पाठशाला MHRD Govt. of India

Now $\overline{x} \leq 0$ under the range $\int_{-\infty}^{0} \implies$ also $\overline{x} - 2\mu \leq 0$ as $\mu \geq 0$.

$\implies \overline{x}(\overline{x} - 2\mu) \geq 0 \implies \int_{-\infty}^{0} \overline{x}(\overline{x} - 2\mu) f_{\overline{X}}(\overline{x}) d\overline{x} \geq 0.$

Hence justified.

Note. If $\mu \in (0, \infty)$ in the above illustration, the asymptotic distribution of $\hat{\mu}$ and $\overline{X}$ is same. For this reason one can use $\overline{X}$ as the MLE of μ in this situation.

TRY YOURSELF!

M2.1. Based on a random sample of size n from a normal $(0, \theta)$ distribution with $c \leq \theta \leq 2c$; c(known), find the MLE of θ.

M2.2. A random sample of size n be drawn from a population with density

$$
\begin{aligned}
f_\theta(x) &= \frac{\theta}{x^{\theta+1}} \quad 1 < x < \infty \\
&= 0 \quad \text{o.w.}
\end{aligned}
$$

Find MLE of $\theta \in [K, \infty)$, if exists where K is a given positive number.

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 2 . . .

M2.1. The likelihood function of θ for given $x_1, \ldots, x_n$ is

$$L(\theta) = \frac{1}{(2\pi\,\theta)^{n/2}} \exp\left\{ -\frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{\theta} \right\}$$

$\Longrightarrow$ the loglikelihood of θ is

$$l_x(\theta) = \text{ constant } - \frac{n}{2}\log\theta - \frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{\theta}$$

e-pathshala पाठशाला MHRD Govt. of India

So

$$l'_x(\theta) = -\frac{n}{2\theta} + \frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{\theta^2} = -\frac{n}{2\theta} + \frac{1}{2}\frac{ns_0^2}{\theta^2}$$

where $s_0^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Case I. $c \leq s_0^2 \leq 2c$

Here both s_0^2 and $\theta \in [c, 2c]$. Then

$l'_x(\theta) = 0 \implies \hat{\theta} = s_0^2$ and

$l''_x(\theta)|_{\theta = s_0^2} = -\frac{n}{2{s_0^2}^2}$

$\implies$ MLE of θ under case I is S_0^2

e-pathshala A Gateway to All Post Graduate Courses

Case II. $\leq s_0^2 < c$

Here $s_0^2 - \theta < 0$

$\implies$ $l_x'(\theta) = \frac{n}{2\theta^2}(s_0^2 - \theta) < 0$

$\implies$ $L(\theta) \downarrow \theta$

$\implies$ MLE of θ under this case is c.

Case III. $\leq s_0^2 > 2c$

Here $s_0^2 - \theta > 0$

$\implies l_x'(\theta) > 0$

$\implies L(\theta) \uparrow \theta$

$\implies$ MLE of θ under case III is $2c$.

MLE on Truncated Parameter Space

Hence combining all these cases

$$
\begin{aligned}
\text{MLE } \hat{\theta} &= c \text{ if } s_0^2 < c \\
&= s_0^2 \text{ if } c \le s_0^2 \le 2c \\
&= 2c \text{ if } s_0^2 > 2c
\end{aligned}
$$

Learning Objectives

- MLE may be biased
- MLE may not be unique
- MLE may not exist
- MLE may be worthless
- MLE may not be sufficient, consistent, asymptotically normal

SOME FEATURES OF ML ESTIMATORS : MLE may be biased :

Ex. Let $X_1, X_2, \ldots, X_n$ be *iid* $R(0, \theta)$ variables; $\theta \in (0, \infty)$. Then, given $\mathbf{x}$, the likelihood function is

$$\begin{aligned} L_{\mathbf{x}}(\theta) &= \frac{1}{\theta^n}, 0 < x_i \leq \theta; i = 1, 2, \ldots, n \Leftrightarrow \theta \geq x_{(n)} \\ &= 0, \quad \text{otherwise} \end{aligned}$$

where $x_{(n)} = \max \ \{x_1, x_2, \ldots, x_n\}$.

So $\hat{\theta}_{MLE} = X_{(n)}$, the maximum likelihood estimator of θ. (Already discussed)

SOME FEATURES OF ML ESTIMATORS : MLE may be biased :

The density of $X_{(n)}$ is

$$\begin{aligned} g(t) &= \frac{n}{\theta^n} t^{n-1} \; 0 < t \leq \theta \\ &= 0 \text{ otherwise} \end{aligned}$$

Now check that $E(X_{(n)}) = \frac{n}{n+1}\theta$ i.e. $\neq \theta$ but $\rightarrow \theta$ as $n \rightarrow \infty$.

So $X_{(n)}$ is biased and only asymptotically unbiased for θ.

SOME FEATURES OF ML ESTIMATORS : MLE may be biased :

Ex. Let $X_1, X_2, \ldots, X_n$ be a random sample from

$N(\theta_1, \theta_2)$; $-\infty < \theta_1 < \infty$, $\theta_2 > 0$

Then the MLE of θ_2 is $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, the sample variance with divisor n (To be discussed later).

Here S^2 is not an unbiased estimator of θ_2 because

$$E(S^2) = \frac{n-1}{n}\ \theta_2$$

Note. S^2 is asymptotically unbiased for θ_2 as under $n \longrightarrow \infty$

$$E(S^2) \longrightarrow \ \theta_2.$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be unique:

- **Ex.** Let $X_1, X_2, \ldots, X_n$ be *iid* $R(\theta, \theta+1)$ variables; $\theta \in (-\infty, \infty)$. Then

$$\begin{aligned} L_{\mathbf{x}}(\theta) &= 1, \quad \theta \le x_i \le \theta+1; i = 1, 2, \ldots, n \\ &= 0, \quad \text{otherwise} \end{aligned}$$

That is

$$\begin{aligned} L_{\mathbf{x}}(\theta) &= 1, \quad x_{(n)} - 1 \le \theta \le x_{(1)} \\ &= 0, \quad \text{otherwise} \end{aligned}$$

$\Longrightarrow$ any value in $\left[x_{(n)} - 1, x_{(1)}\right]$ is a maximum likelihood estimate of θ.

Note. If $x_{(n)} - 1 = x_{(1)}$; then unique *ml* estimate of θ is $x_{(1)} = x_{(n)} - 1$.

SOME FEATURES OF ML ESTIMATORS : MLE may not be unique:

Ex. Let $X \sim$ pmf$f$; $f \in \{f_1, f_2, f_3, f_4\}$ where

$$\begin{aligned} f_1(x) &\equiv \text{Bin } (5, 0.8) \\ f_2(x) &\equiv \text{Poisson } (1) \\ f_3(x) &= (0.8)(0.2)^x \\ f_4(x) &= -\frac{(0.4)^{x+1}}{(x+1)\ln(0.6)} \end{aligned}$$

Given $X = 0$ find the most likely choice of f.

Here f is the unknown parameter θ. Given $X = 0$ the likelihood function $L(\theta|X = 0) = P[X = 0|\theta]$. Hence

$$\begin{aligned} L(\theta|X = 0) &= 0.8 \ , \ \ \theta = f_1 \\ &= 0.27 \ , \ \ \theta = f_2 \end{aligned}$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be unique:

Ex. Let $X \sim$ pmf$f$; $f \in \{f_1, f_2, f_3, f_4\}$ where

$$
\begin{aligned}
f_1(x) &\equiv \text{Bin } (5, 0.8) \\
f_2(x) &\equiv \text{Poisson } (1) \\
f_3(x) &= (0.8)(0.2)^x \\
f_4(x) &= -\frac{(0.4)^{x+1}}{(x+1)\ln(0.6)}
\end{aligned}
$$

Given $X = 0$ find the most likely choice of f.

Here f is the unknown parameter θ. Given $X = 0$ the likelihood function $L(\theta|X = 0) = P[X = 0|\theta]$. Hence

$$
\begin{aligned}
L(\theta|X = 0) &= 0.8 \;, \quad \theta = f_1 \\
&= 0.27 \;, \quad \theta = f_2
\end{aligned}
$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be unique:

$$\begin{aligned} L(\theta|X=0) &= 0.8 \ , \ \ \theta = f_3 \\ &= 0.78 \ , \ \ \theta = f_4 \end{aligned}$$

So $L(\theta)$ is maximum for any of f_1 and f_3.

$\Longrightarrow \ X = 0$ MLE of $\theta \in \{f_1, f_3\}$.

Clearly MLE is not unique in this case.

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

- **Ex.** Let $X_1, X_2, \ldots, X_n$ be *iid* Bernoulli (p) variables; $p \in (0, 1)$. Here the likelihood function of p is

$$L_{\mathbf{x}}(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- If $\sum x_i = 0 \Leftrightarrow x_i = 0, i = 1, 2, \ldots, n$ then $L_{\mathbf{x}}(p) = (1-p)^n$ which is supremum at $p = 0$; but $p > 0$.
- If $\sum x_i = n \Leftrightarrow x_i = 1, i = 1, 2, \ldots, n$ then $L_{\mathbf{x}}(p) = p^n$ which is supremum at $p = 1$; but $p < 1$.

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

Thus in each of the two extreme cases (i.e. $\bar{x} = 0$ or 1) the supremum is attained at the boundary of $\Omega = (0, 1)$.

$$0 < \sum x_i < n \implies L_{\mathbf{x}}(p) \text{ is supremum at } p = \frac{\sum x_i}{n} \in (0, 1).$$

$$\implies \overline{X} = \frac{\sum x_i}{n} \text{ is the MLE of } p, \text{ provided } 0 < \sum x_i < n; \text{ i.e. } 0 < \bar{x} < 1;$$

Otherwise MLE of p does not exist.

Note1: If $0 \leq p \leq 1$; $\hat{p}_{MLE} = \overline{X}$

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

Note2:$p \in (0,1) \implies$ some trouble in defining the MLE of p if the samples $\Rightarrow$ either $\overline{x} = 0$ or 1. But

$$P[\overline{X} = 0 \text{ or } 1] = P[\sum_{i=1}^{n} X_i = 0 \text{ or } n] = (1-p)^n + p^n \longrightarrow 0.$$

as $n \longrightarrow \infty$ for $p \in (0,1)$.

$\implies$ the probability of occuring such an extreme case is negligible if n is large enough.

$\implies$ for large n $\overline{X}$ can be used as the MLE of $p \in (0,1)$ in practice.

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

Note2:$p \in (0,1) \implies$ some trouble in defining the MLE of p if the samples $\Rightarrow$ either $\overline{x} = 0$ or 1. But

$$P[\overline{X} = 0 \text{ or } 1] = P[\sum_{i=1}^{n} X_i = 0 \text{ or } n] = (1-p)^n + p^n \longrightarrow 0.$$

as $n \longrightarrow \infty$ for $p \in (0,1)$.

$\implies$ the probability of occuring such an extreme case is negligible if n is large enough.

$\implies$ for large n $\overline{X}$ can be used as the MLE of $p \in (0,1)$ in practice.

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

An alternative way to remove the undefinedness of MLE if $p \in (0,1)$

Take MLE as

$$
\begin{aligned}
\hat{p} &= \epsilon^{*} \text{ if } \overline{x} = 0 \\
&= \overline{x} \text{ if } 0 < \overline{x} < 1 \\
&= 1 - \epsilon^{**} \text{ if } \overline{x} = 1
\end{aligned}
$$

where ϵ^{*} and ϵ^{**} are infinitesimally small positive numbers.

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

Obviously

$$\sqrt{n}(\hat{p} - p) = \sqrt{n}(\overline{X} - p) + \sqrt{n}(\hat{p} - \overline{X})$$

From CLT $\sqrt{n}(\overline{X} - p) \xrightarrow{\mathcal{D}} N(0, p(1-p))$

Also as $\hat{p} - \overline{X} = 0 \Longrightarrow \{|\hat{p} - \overline{X}| < \epsilon\}$ for every $\epsilon > 0$, so

$P[\sqrt{n}|\hat{p} - \overline{X}| < \epsilon] \geq P[\hat{p} = \overline{X}] = P[0 < \overline{X} < 1] = 1 - (1-p)^n - p^n \longrightarrow 1$

as $n \longrightarrow \infty$ for every $\epsilon > 0$ for $p \in (0, 1)$.

$\Longrightarrow \hat{p} - \overline{X} \xrightarrow{P} 0$

SOME FEATURES OF ML ESTIMATORS : MLE may not exist:

Obviously

$$\sqrt{n}(\hat{p} - p) = \sqrt{n}(\overline{X} - p) + \sqrt{n}(\hat{p} - \overline{X})$$

From CLT $\sqrt{n}(\overline{X} - p) \xrightarrow{\mathcal{D}} N(0, p(1 - p))$

Also as $\hat{p} - \overline{X} = 0 \Longrightarrow \{|\hat{p} - \overline{X}| < \epsilon\}$ for every $\epsilon > 0$, so

$P[\sqrt{n}|\hat{p}-\overline{X}| < \epsilon] \geq P[\hat{p} = \overline{X}] = P[0 < \overline{X} < 1] = 1-(1-p)^n - p^n \longrightarrow 1$

as $n \longrightarrow \infty$ for every $\epsilon > 0$ for $p \in (0, 1)$.

$\Longrightarrow \hat{p} - \overline{X} \xrightarrow{P} 0$

SOME FEATURES OF ML ESTIMATORS : Some Features of ML Estimators :

So, by Slutsky's theorem (i.e. $X \xrightarrow{P} c$, $Y \xrightarrow{\mathcal{D}} Z \implies X + Y \xrightarrow{\mathcal{D}} Z + c$)
we get

$$\sqrt{n}(\hat{p} - p) \xrightarrow{\mathcal{D}} N(0, p(1-p))$$

So the proposed MLE $\hat{p}$ and $\overline{X}$ are equivalent at least asymptotically.

SOME FEATURES OF ML ESTIMATORS : MLE may be worthless:

- **Ex.** Let $X \sim Bernoulli(p)$; $\frac{1}{4} \leq p \leq \frac{3}{4}$. Then

$$\begin{aligned} L_x(p) &= 1 - p \ ; \ x = 0 \\ &= p \ ; \qquad x = 1 \end{aligned}$$

$\Longrightarrow$ given $x = 0$; $\hat{p}_{MLE} = \frac{1}{4}$ and for $x = 1$; $\hat{p}_{MLE} = \frac{3}{4}$.

Only two co-ordinate points $(0, \frac{1}{4})$ and $(1, \frac{3}{4})$ on $(X, \hat{p}_{MLE}(X)) \Rightarrow \hat{p}_{MLE} = \frac{2X+1}{4}$.

$\Big[$**other possibilites are:** $\frac{x\,2^X+1}{4}$, $\frac{3^X}{4}$ **etc which are all equivalent**$\Big]$

SOME FEATURES OF ML ESTIMATORS : MLE may be worthless:

$\Longrightarrow$

$$\begin{aligned} MSE(\hat{p}_{MLE}) &= E_p\left(\hat{p}_{MLE} - p\right)^2 \\ &= E_p\left(\frac{2X+1}{4} - p\right)^2 \\ &= \frac{1}{16} \quad \text{(on simplification)} \end{aligned}$$

Now, just take a trivial estimator of p, say $\hat{p^*} = \frac{1}{2}$

MSE $(\hat{p^*}) = (\frac{1}{2} - p)^2 \leq \frac{1}{16}$ $\forall p \in [\frac{1}{4}, \frac{3}{4}]$ with $<$ for $p \neq \frac{1}{4}, \frac{3}{4}$.

$\Longrightarrow$ performance of $\hat{p}_{MLE}$ is uniformly worse even than $\hat{p^*}$, a trivial estimator of p.

SOME FEATURES OF ML ESTIMATORS : MLE may not be consistent:

- **Ex.** Let $X_{ij} \sim N(\mu_i, \sigma^2)$; $i = 1, 2, \ldots, n$, $j = 1, 2$

$$L_{\mathbf{x}}(\theta) = \frac{1}{(2\pi)^n \sigma^{2n}} \exp\left[-\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij} - \mu_i)^2/2\sigma^2\right]$$

- The MLE of μ_i and σ^2 are respectively $\hat{\mu}_i = \overline{x}_i$ and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij}-\overline{x}_i)^2}{2n}$$ (*To be discussed later*).

SOME FEATURES OF ML ESTIMATORS : MLE may not be consistent:

- On simplification, finally we have

$$\hat{\sigma^2} \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i1}-x_{i2})^2}{2\sigma^2}\frac{\sigma^2}{2} \xrightarrow{P} \frac{\sigma^2}{2}$$

(as by Khinchin's WLLN $\frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i1}-x_{i2})^2}{2\sigma^2} \xrightarrow{P} 1$)

- Hence $\hat{\sigma^2} \xrightarrow{P} \frac{\sigma^2}{2} \implies \hat{\sigma^2}$ is not consistent for σ^2. [**Neyman and Scott**]

SOME FEATURES OF ML ESTIMATORS : MLE may not be asymptotically normal:

- **Ex.** Let $X_1, X_2, \ldots, X_n$ be *iid* $R(0, \theta)$ variables; $\theta \in (0, \infty)$. Then $\hat{\theta}_{MLE} = X_{(n)}$, the maximum order statistic. (*Try yourself*)
-

$$
\begin{aligned}
P[n(\theta - X_{(n)}) \leq t] &= 1 - P[X_{(n)} < \theta - t/n] \\
&= 1 - \frac{n}{\theta^n} \int_0^{\theta - t/n} t^{n-1} dt \\
&= 1 - \left(1 - \frac{t}{n\theta}\right)^n \longrightarrow 1 - \exp(-t/\theta)
\end{aligned}
$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be asymptotically normal:

- $\Longrightarrow \; n(\theta - X_{(n)}) \to Z \sim$ exponential distribution with density $\frac{1}{\theta}\exp(-x/\theta)$; as $n \to \infty$.
- **Note:** Here $\sqrt{n}(\theta - X_{(n)}) = \frac{n(\theta - X_{(n)})}{\sqrt{n}} \xrightarrow{\mathcal{D}} 0$.

SOME FEATURES OF ML ESTIMATORS : MLE may not be asymptotically normal:

Ex. Let $X_1, \ldots, X_n$ follow independently the common positive density

$$f_\theta(x) = \exp\{-(x-\theta)\} \ , \ x \geq \theta \ ; \ \theta \in (-\infty, \infty).$$

Find the asymptotic distribution of $X_{(1)} - \theta$, where
$X_{(1)} = \min\{X_1, \ldots, X_n\}$.

The distribution function of $Y_n = X_{(1)} - \theta$ is

$$\begin{aligned} F_n(y) &= 0 \text{ if } y \leq 0 \\ &= 1 - \exp(-ny) \text{ if } y > 0 \end{aligned}$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be asymptotically normal:

Ex. Let $X_1, \ldots, X_n$ follow independently the common positive density

$$f_\theta(x) = \exp\{-(x-\theta)\} \ , \ x \geq \theta \ ; \ \theta \in (-\infty, \infty).$$

Find the asymptotic distribution of $X_{(1)} - \theta$, where $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

The distribution function of $Y_n = X_{(1)} - \theta$ is

$$\begin{aligned} F_n(y) &= 0 \text{ if } y \leq 0 \\ &= 1 - \exp(-ny) \text{ if } y > 0 \end{aligned}$$

SOME FEATURES OF ML ESTIMATORS : MLE may not be asymptotically normal:

Define

$$F^*(y) = 0 \text{ if } y < 0$$
$$= 1 \text{ if } y \geq 0$$

Check that

(i) $F^*(y)$ is a distribution function,

(ii) $F^*(y)$ is the distribution function of a degenerate probability distribution with degeneracy at $y = 0$, and

(iii) $F_n(y) \longrightarrow F^*(y)$ for all y except at $y = 0$, the only point of discontinuity.

Invariance Property and Likelihood Equation of MLE

Module 4

Saurav De

Department of Statistics
Presidency University

Invariance Property and Likelihood Equation of MLE

MLE and Invariance Property

Let $\hat{\theta}$ be MLE of θ. Then for the parametric function $g(\theta) : \Omega \to \Gamma$; MLE is $g(\hat{\theta})$.

Proof. Let us define $\Omega_\gamma = \{\theta : g(\theta) = \gamma\}$. This means $\Omega = \bigcup_{\gamma \in \Gamma} \Omega_\gamma$.

Again let $M_x(\gamma) = \sup_{\theta \in \Omega_\gamma} L_x(\theta) =$ Likelihood function induced by g.

We are to find $\hat{\gamma}$ at which $M_x(\gamma)$ is maximised.

Invariance Property and Likelihood Equation of MLE

e-Pathshala पाठशाला MHRD Govt. of India

Now $M_x(\hat{\gamma}_0) = \sup_{\theta \in \Omega_{\hat{\gamma}_0}} L_x(\theta) \geq \underset{\theta \in \Omega_{\hat{\gamma}_0}}{L_x(\hat{\theta})}$

where $\Omega_{\hat{\gamma}_0} = \{\theta : g(\theta) = \hat{\gamma}_0\}$. As $g(\hat{\theta}) = \hat{\gamma}_0$ so $\hat{\theta} \in \Omega_{\hat{\gamma}_0}$

Again

$$
\begin{aligned}
M_x(\hat{\gamma}_0) \leq \sup_{\hat{\gamma} \in \Gamma} M_x(\hat{\gamma}) &= \sup_{\hat{\gamma} \in \Omega} \sup_{\theta \in \Omega_{\hat{\gamma}_0}} L_x(\theta) \\
&= \sup_{\theta \in \Omega} L_x(\theta) = L_x(\hat{\theta})
\end{aligned}
$$

Invariance Property and Likelihood Equation of MLE

Therefore

$$M_x(\hat{\gamma}_0) = L_x(\hat{\theta}) = \sup_{\gamma \in \Gamma} M_x(\gamma).$$

Hence $\hat{\gamma}_0$ is the MLE of γ, i.e. $g(\hat{\theta})(= \hat{\gamma}_0)$ is the MLE of $g(\theta)(= \gamma)$.
Proved

Ex.3 Let $X_1, X_2, \ldots, X_n \sim \text{Bin}\ (1, p)\ \ ;\ \ 0 \leq p \leq 1$

Then $V_p(X) = p(1-p)(= g(p))$ and $\hat{p}_{MLE} = \overline{X}_n$.

By invariance property MLE of $V_p(X)$ is
$g(\hat{p}_{MLE}) = \hat{p}_{MLE}(1 - \hat{p}_{MLE}) = \overline{X}_n(1 - \overline{X}_n)$.

Invariance Property and Likelihood Equation of MLE

Application 4. Suppose that n observations are taken on a random variable X with distribution $N(\mu, 1)$, but instead of recording all the observations, one notes only whether or not the observation is less than 0. If $\{X < 0\}$ occurs $m(< n)$ times, find the MLE of μ.

Let $X_1, X_2, \ldots, X_n \sim N(\mu, 1)$.

Let $\theta = P_\mu[X_1 < 0] = \Phi(-\mu) = 1 - \Phi(\mu)$.

This means $\mu = -\Phi^{-1}(\theta)$, a continuous function of θ.

Invariance Property and Likelihood Equation of MLE

$$\begin{aligned} Y_i &= 1 \text{ if } X_i < 0 \\ &= 0 \text{ if } X_i \geq 0 \end{aligned}$$

Then $Y_1, Y_2, \ldots, Y_n \sim \text{Bin}\,(1, \theta)\ ;\ \ 0 \leq \theta \leq 1,$

Now the MLE of θ is $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{\#\{X_i<0\}}{n} = \frac{m}{n}$. (*See the Application 3*).

Hence by the invariance property the MLE of μ is $-\Phi^{-1}(\frac{m}{n})$.

Invariance Property and Likelihood Equation of MLE

Application. A commuter trip consists of first riding a subway to the bus stop and then taking a bus. The bus she should like to catch arrives uniformly over the interval θ_1 to θ_2. She would like to estimate both θ_1 and θ_2 so that she would have some idea about the time she should be at the bus stop (θ_1) and she should have too late and wait for the next bus (θ_2). Over an 8-day period she makes certain to be at the bus stop so early not to miss the bus and records the following arrival time of the bus.

5:15 PM , 5:21 PM , 5:14 PM , 5:23 PM , 5:29 PM , 5:17 PM , 5:15 PM , 5:18 PM

Estimate θ_1 and θ_2. Also give the MLEs for the mean and the variability of the arrival distribution.

Invariance Property and Likelihood Equation of MLE

Solution: Let T : arrival time of the bus. Then according to the question $T \sim R(\theta_1, \theta_2)$.

Let $T_1, T_2, \ldots, T_n$ be n independent random arrival times of the bus.

Then the likelihood of $\theta = (\theta_1, \theta_2)$ is

$$\begin{aligned} L(\theta) &= \frac{1}{(\theta_2 - \theta_1)^n} \text{ if } \theta_1 \leq t_i \leq \theta_2; i = 1, 2, \ldots, n \\ &= 0 \text{ otherwise} \end{aligned}$$

$$\begin{aligned} \text{i.e. } L(\theta) &= \frac{1}{(\theta_2 - \theta_1)^n} \text{ if } \theta_1 \leq t_{(1)} < t_{(n)} \leq \theta_2 \\ &= 0 \text{ otherwise} \end{aligned}$$

where $t_{(1)} = \min\{t_1, t_2, \ldots, t_n\}$ and $t_{(n)} = \max\{t_1, t_2, \ldots, t_n\}$.

Invariance Property and Likelihood Equation of MLE

Under $\theta_1 \leq t_{(1)} < t_{(n)} \leq \theta_2$, $t_{(n)} - t_{(1)} \leq \theta_2 - \theta_1$.

Hence $L(\theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} \leq \frac{1}{(t_{(n)} - t_{(1)})^n} = L(t_{(1)}, t_{(n)})$.

Thus MLE of (θ_1, θ_2) is $(T_{(1)}, T_{(n)})$, where $T_{(1)}$ and $T_{(n)}$ are the minimum and maximum order statistic respectively.

Now $E(T) = \frac{\theta_1 + \theta_2}{2}$ and $V(T) = \frac{(\theta_2 - \theta_1)^2}{12}$ are two continuous functions of (θ_1, θ_2).

So MLE of Mean : $\frac{T_{(1)} + T_{(n)}}{2}$ and of variability : $\frac{(T_{(n)} - T_{(1)})}{\sqrt{12}}$ (*from invariance property of MLE*)

Invariance Property and Likelihood Equation of MLE

Computation using R :

(The minute component of the time has been considered as the data)

R Code and Output :

```
> samp=c(15,21,14,23,29,17,15,18)# given sample
> n=length(samp)# size of the sample
> n
[1] 8
> MLE_theta1=min(samp)# MLE of the parameters
> MLE_theta2=max(samp)
> MLE_theta1
[1] 14
> MLE_theta2
[1] 29
> MLE_Mean=(MLE_theta1+MLE_theta2)/2# MLE of mean
> cat("The mean arrival time is :",MLE_Mean,"minutes after 5pm.\n")
The mean arrival time is : 21.5 minutes after 5pm.
> MLE_Var=(MLE_theta2-MLE_theta1)/sqrt(12)# MLE of variability
> MLE_Var
[1] 4.330127
```

Invariance Property and Likelihood Equation of MLE

Likelihood Equations and Related Discussions

$l_x(\theta) = \ln L_x(\theta)$ is called log-likelihood function of θ.

Likelihood Equation : $\frac{\partial l_x(\theta)}{\partial \theta} = 0$.

Any MLE is a root of the likelihood equation.

Any root may be *local minima* or *local maxima.*

Possible verification for the root $\hat{\theta}$ to be an MLE: $\frac{\partial^2 l_x(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$.

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)'$, the likelihood equations: $\frac{\partial l_x(\boldsymbol{\theta})}{\partial \theta_i} = 0 \quad i = 1, \ldots, s$.

Possible verification for the root $\hat{\boldsymbol{\theta}}$ to be an MLE of $\boldsymbol{\theta}$: The Hessian matrix $\left(\left(\frac{\partial^2 l_x(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}|_{\theta_i=\hat{\theta}_i, \theta_j=\hat{\theta}_j}\right)\right)$ is negative definite.

Invariance Property and Likelihood Equation of MLE

Result: Let T be a sufficient statistic for the family of distributions $\{f_\theta : \theta \in \Omega\}$. If a unique MLE of θ exists, it is a (nonconstant) function of T. If a MLE of θ exists but is not unique, one can find a MLE that is a function of T.

Proof. Since T is sufficient, from Neyman-Fisher factorisability we can write

$$L(\theta) = f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x})$$

for all $\mathbf{x}$, all θ and some h and g_θ, where $L(\theta)$ is the likelihood function of θ and $f_\theta(\mathbf{x})$ is the joint pmf or pdf of sample observations $\mathbf{x} = (x_1, \ldots, x_n)$.

Invariance Property and Likelihood Equation of MLE

If $L(\theta)$ is maximised by a unique MLE $\hat{\theta}$, naturally it will also maximise the function $g_\theta(T(\mathbf{x}))$.

$\implies$ $\hat{\theta}$ should be a function of T.

If MLE of θ exists but is not unique, then $\exists$ some MLE θ that can be expressed as a function of T.

Proved

Invariance Property and Likelihood Equation of MLE

Result: Under regular estimation case (*i.e. the situation where all the regularity conditions of Cramer-Rao Inequality hold*) if an estimator $\hat{\theta}$ of θ attains the Cramer-Rao Lower Bound **CRLB** for the variance, the likelihood equation has a unique solution $\hat{\theta}$ that maximises the likelihood function.

Proof. Let $L(\theta|\mathbf{x})$ denote the likelihood function of real-valued parameter θ given the sample observations $\mathbf{x} = (x_1, \ldots, x_n)$. If $f_\theta(\mathbf{x})$ denotes the joint

pmf or pdf of $\mathbf{x}$, from the equality condition in CR inequality we get

$$\frac{\partial}{\partial \theta} \log f_\theta(\mathbf{x}) = k(\theta)(\hat{\theta}(\mathbf{x}) - \theta)$$

Invariance Property and Likelihood Equation of MLE

that is

$$\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) = k(\theta)(\hat{\theta}(\mathbf{x}) - \theta) \qquad (*)$$

with probability 1.

$\Longrightarrow$ the likelihood equation $\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) = 0$ has the unique solution $\theta = \hat{\theta}$.

Differentiating both sides of $(*)$ again with respect to θ we get

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{x}) = k'(\theta)(\hat{\theta} - \theta) - k(\theta).$$

Hence

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{x})|_{\theta=\hat{\theta}} = -k(\theta) \qquad (**)$$

Invariance Property and Likelihood Equation of MLE

Now, If T is an unbiased estimator of a real-valued estimable parametric function $g(\theta)$ which is differentiable at least once, from CR regularity conditions we directly get

$$\int T(\mathbf{x}))f_\theta(\mathbf{x})d\mathbf{x} = g(\theta)$$

Or, differentiating both sides with respect to θ we get

$$\int T(\mathbf{x}))\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})d\mathbf{x} = g'(\theta)$$

In particular choosing $g(\theta) = \theta$ and noting that $E_\theta\left[\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right] = 0$ we get

$$E_\theta\left[(T(\mathbf{X}) - \theta)\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right] = 1.$$

Invariance Property and Likelihood Equation of MLE

Finally substituting

$$T(\mathbf{X}) - \theta = [k(\theta)]^{-1} \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X})$$

(*looking at* $(*)$ *as* $T(\mathbf{X})$ *is just like* $\hat{\theta}(\mathbf{X})$ by its definition) we get

$$[k(\theta)]^{-1} E_\theta \left[\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X}) \right]^2 = 1$$

That is

$$k(\theta) = E_\theta \left[\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X}) \right]^2 > 0 \text{ by regularity condition}$$

Thus from $(**)$ the S.O.C. for maximising $L(\theta)$ holds. Hence proved.

Invariance Property and Likelihood Equation of MLE

Result: Suppose $\frac{\partial^2 l_x(\theta)}{\partial\theta^2} \leq 0 \;\forall\; \theta \in \Omega$. Then $\hat{\theta}$ satisfying $\frac{\partial l_x(\theta)}{\partial\theta} = 0$ is the global maxima.

Proof. $l_x(\theta) = l_x(\hat{\theta}) + (\theta - \hat{\theta})\frac{\partial l_x(\theta)}{\partial\theta}|_{\theta=\hat{\theta}} + \frac{(\theta-\hat{\theta})^2}{2}\frac{\partial^2 l_x(\theta)}{\partial\theta^2}|_{\theta=\theta^*} \quad \theta^* \in (\hat{\theta}, \theta)$.

Note that the RHS $\leq l_x(\hat{\theta})$ because, in RHS the 2nd factor of 2nd term vanishes and the 2nd factor of the third term ≤ 0. Hence proved.

Invariance Property and Likelihood Equation of MLE

Result: Suppose

(i) $\frac{\partial l_x(\theta)}{\partial \theta} = 0$ iff $\theta = \hat{\theta}$.

(ii) $\frac{\partial^2 l_x(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$. And

(iii) $\hat{\theta}$ is an interior point of an interval $I \subset \Omega$.

Then $\hat{\theta}$ is the global maxima.

Proof. If possible suppose $\hat{\theta}^*$ is such that $l_x(\hat{\theta}^*) > l_x(\hat{\theta})$. Then there must be a local minima in between the local maxima $\hat{\theta}$ and $\hat{\theta}^*$. This means for that minima point also $\frac{\partial l_x(\theta)}{\partial \theta} = 0$, which is a contradiction to the supposition (i). Hence proved.

Invaraince Property and Likelihood Equation of MLE

Ex.1 Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ independently.

Then $l_{\mathbf{x}}(\theta) =$ constant - $\frac{\sum (x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} ln\sigma^2$.

So $\frac{\partial l_{\mathbf{x}}(\theta)}{\partial \mu} = 0$ and $\frac{\partial l_{\mathbf{x}}(\theta)}{\partial \sigma^2} = 0$ imply $\hat{\mu} = \overline{x}$, $\hat{\sigma^2} = \frac{1}{n} \sum (x_i - \overline{x})^2 = s^2$.

Also check that $\frac{\partial^2 l_{\mathbf{x}}(\theta)}{\partial \mu \partial \sigma^2}|_{(\mu, \sigma^2) = (\overline{x}, s^2)} = 0$,

Invariance Property and Likelihood Equation of MLE

$$\frac{\partial^2 l_{\mathbf{x}}(\theta)}{\partial\mu^2}\Big|_{(\mu,\sigma^2)=(\overline{x},s^2)} = -\frac{n}{s^2} \quad , \quad \frac{\partial^2 l_{\mathbf{x}}(\theta)}{\partial\partial(\sigma^2)^2}\Big|_{(\mu,\sigma^2)=(\overline{x},s^2)} = -\frac{n}{s^4}.$$

So the Hessian matrix

$$H = \begin{pmatrix} -n/s^2 & 0 \\ 0 & -n/2s^4 \end{pmatrix}$$

is negative definite. Hence $(\overline{X}, S^2)$ is the global maxima point and is the MLE of (μ, σ^2).

Invariance Property and Likelihood Equation of MLE

Aliter: If $\psi(x) = x - 1 - lnx$ then
$\psi'(x) = 1 - 1/x$, $\psi''(x) = 1/x^2 > 0$.

Therefore $\psi(x)$ is minimum at $x = 1$ and min$\psi(x) = 1 - 1 - 0 = 0$. Based on this result we can write

$$l_{\mathbf{x}}(\hat{\mu}, \hat{\sigma^2}) - l_{\mathbf{x}}(\mu, \sigma^2) = \frac{\sum (x_i - \mu)^2}{2\sigma^2} + \frac{n}{2} ln\sigma^2 - \frac{n}{2} - \frac{n}{2} lns^2$$

$$\geq \frac{ns^2}{2\sigma^2} - \frac{n}{2} \ln \frac{s^2}{\sigma^2} - \frac{n}{2} = \frac{n}{2}\left[\frac{s^2}{\sigma^2} - 1 - \ln \frac{s^2}{\sigma^2}\right] \geq 0.$$

Invariance Property and Likelihood Equation of MLE

TRY YOURSELF!

M4.1. Let $X_1, X_2, \ldots, X_n \sim \text{Lognormal}(\mu, \sigma^2)$ independently. Then MLE of (μ, σ^2) is $(\overline{Y}, S^{*2})$ where $Y = \log X$ and $(\overline{Y}, S^{*2}) =$ (sample mean , sample variance(with divisor n)) on Y.

Hint. If $X \sim \text{Lognormal}(\mu, \sigma^2)$ then $Y = \log X \sim N(\mu, \sigma^2)$. Now proceed as in Ex. 1.

M4.2. (*continuation*)If in **M4.1.** $\mu = 0$, find the MLE of σ^2

M4.3. Consider a random sample of size n from Exponential (mean $= \beta$). It is given only that k, $0 < k < n$, of these n observations are $\leq M$, where M is a known positive number. Find the MLE of β.

Invariance Property and Likelihood Equation of MLE

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 4 . . .

M4.2. If $X \sim \text{Lognormal}(0, \sigma^2)$ then $Y = \log X \sim N(0, \sigma^2)$.

Given $\mathbf{y} = (y_1, \ldots, y_n)$, the loglikelihood function of σ^2 is

$$l(\sigma^2) = \text{ constant } - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2$$

Now using maxima-minima principle from F.O.C. we get

$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \hat{\sigma}^2$ (say) is the only solution of the likelihood equation.

Invariance Property and Likelihood Equation of MLE

Also $\frac{\partial^2}{\partial(\sigma^2)^2} l(\sigma^2)|_{\sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} < 0.$

Moreover $\hat{\sigma}^2$ is an interior point of the parameterspace $\Omega = (0, \infty)$.

$\Longrightarrow$ $\hat{\sigma}^2$ is the point of global maxima of the likelihood function i.e. the unique MLE of σ^2.

Invariance Property and Likelihood Equation of MLE

M4.3. Let $X_1, X_2, \ldots, X_n \sim$ Exponential (mean $= \beta$).

Define $Y_i = 1(0)$ if $X_i \leq M$ $(X_i > M)$.

Then $Y_1, Y_2, \ldots, Y_n \sim$ Bin $(1, \theta)$; $0 < \theta < 1$,

where $\theta = P_\beta[X_1 \leq M] = 1 - \exp\left[-\frac{M}{\beta}\right]$. This means $\beta = -\frac{M}{\log(1-\theta)}$, a continuous function of θ.

Now the MLE of θ is $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{\#\{X_i \leq M\}}{n} = \frac{k}{n}$. (*See the Application 4*).

Hence by the invariance property the MLE of β is $-\frac{M}{\log(1-\frac{k}{n})}$.

Multivariate Multiparameter MLE

Module 5

Saurav De

Department of Statistics
Presidency University

e-Pathshala पाठशाला MHRD Govt. of India

Ex. (*Multivariate multiparameter case*) $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \Sigma)$, a p-variate normal distribution.

The likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ for given $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is

$$L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right]$$

Multivariate Multiparameter MLE

Now the exponent

$$\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_\alpha-\boldsymbol{\mu}) = \sum_{\alpha=1}^{n}(\mathbf{x}_\alpha-\bar{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$$

The second term in RHS is a positive definite (p.d.) quadratic form as Σ is a p.d. matrix. Hence the term will be minimum (i.e. 0) iff $\boldsymbol{\mu} = \bar{\mathbf{x}}$.

Multivariate Multiparameter MLE

$\implies$ keeping Σ to be fixed for the time being, $L(\boldsymbol{\mu}, \Sigma)$ will be maximised at $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$.

Then

$$L(\hat{\boldsymbol{\mu}}, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \bar{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})\right]$$

So the loglikelihood is

$$\ell(\Sigma) = \text{ constant } + \frac{n}{2}\log|\Sigma^{-1} S| - \frac{1}{2}\sum_{\alpha=1}^{n} tr\ (\mathbf{x}_\alpha - \bar{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_\alpha - \bar{\mathbf{x}})$$

[as a scalar is the trace of itself]

where $S = \frac{1}{n}\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \overline{\mathbf{x}})(\mathbf{x}_\alpha - \overline{\mathbf{x}})'$ is the sample variance-covariance matrix with divisor n.

(*In likelihood function data X is given and hence the matrix S also becomes known and is included in Constant term*)

This means

$$\ell(\Sigma) = \text{ constant } + \frac{n}{2}\log|\Sigma^{-1} S| - \frac{1}{2}\, tr\ \Sigma^{-1} \sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \overline{\mathbf{x}})(\mathbf{x}_\alpha - \overline{\mathbf{x}})'$$

(*as $tr\ AB = tr\ BA$*) Or

$$\ell(\Sigma) = \text{ constant } + \frac{n}{2}\log|\Sigma^{-1} S| - \frac{1}{2}\, tr\ n\Sigma^{-1} S$$

i.e.

$$\ell(A) = \text{ constant } + \frac{n}{2}\log|A| - \frac{n}{2}\, tr\ A$$

where $A = \Sigma^{-1} S$.

Multivariate Multiparameter MLE

Let $\lambda_1, \ldots, \lambda_p$ be the p eigen values of A. As both Σ and S are p.d. matrices, so also is A.

$\implies$ all $\lambda_i > 0$. Also we know $|A| = \prod_{i=1}^{p} \lambda_i$ and $tr\ A = \sum_{i=1}^{p} \lambda_i$. Using all these we get

$$\ell = \ell(\boldsymbol{\lambda}) = \ell(A) = \text{ constant } + \frac{n}{2} \sum_{i=1}^{p} \log \lambda_i - \frac{n}{2} \sum_{i=1}^{p} \lambda_i$$

Multivariate Multiparameter MLE

Now

$$\frac{\partial}{\partial \lambda_i}\ell = 0 \ \forall\ i \implies \lambda_i = 1 \ \forall\ i$$

$$\frac{\partial^2}{\partial \lambda_i^2}\ell|_{\lambda_i=1} = -\frac{n}{2} < 0 \ \forall\ i$$

and

$$\frac{\partial^2}{\partial \lambda_i \partial \lambda_j}\ell = 0 \forall\ i,j\ ;\ i \neq j.$$

Hence by Maxima-Minima principle $\lambda_i = 1 \ \forall\ i$ maximise the likelihood function.

Also $\lambda_i = 1 \ \forall\ i \implies A = I$, the identity matrix.

$\implies \Sigma^{-1} S = I \implies \Sigma = S (= \hat{\Sigma})(\text{say})$

That is at $\Sigma = S$ the likelihood is maximised.

Thus the MLE of $(\boldsymbol{\mu}, \Sigma)$ is $(\overline{\mathbf{X}}, S)$.

Multivariate Multiparameter MLE

Application. (*Bivariate Normal Distribution*) This is a particular case of the above example with $p = 2$.

Let $\mathbf{X} = (X_1, X_2)' \sim N_2(\boldsymbol{\mu}, \Sigma)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

and $\rho = Corr(X_1, X_2)$.

A Gateway to All Post Graduate Courses

Based on a random sample of size n on (X_1, X_2) the MLE of $(\boldsymbol{\mu}, \Sigma)$ will be

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \end{pmatrix} \text{ and } S = \frac{1}{n}\sum_{\alpha=1}^{n}(\mathbf{X}_\alpha - \overline{\mathbf{X}})(\mathbf{X}_\alpha - \overline{\mathbf{X}})' = \begin{pmatrix} S_1^2 & S_{12} \\ S_{12} & S_1^2 \end{pmatrix}$$

where $S_1^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(X_{1\alpha} - \overline{X}_1)^2$ is the sample variance on the random variable X_1 etc. and $S_{12} = \frac{1}{n}\sum_{\alpha=1}^{n}(X_{1\alpha} - \overline{X}_1)(X_{2\alpha} - \overline{X}_2)$ is the sample covariance on (X_1, X_2).

Hence MLE of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ is $(\overline{X}_1, \overline{X}_2, S_1^2, S_2^2, r)$ where $r = \frac{S_{12}}{\sqrt{S_1^2 S_2^2}}$ = sample correlation coefficient of (X_1, X_2).

A Gateway to All Post Graduate Courses

Application. (*Bivariate Normal Distribution with reparameterization technique*)

Let (X, Y) have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$. Suppose that n observations are made on (X, Y), and $N - n$ observations on X only; i.e. $N - n$ observations on Y are missing. Find the MLEs of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$.

We know that for $(X, Y) \sim$ BVN $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$,

$$X \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y|X = x \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)).$$

W.l.g. denote $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ by θ for simplicity.

Multivariate Multiparameter MLE

Define $\alpha = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1$, $\beta = \rho\frac{\sigma_2}{\sigma_1}$ and $\gamma^2 = \sigma_2^2(1-\rho^2)$ . . . . . . $(*)$
$\implies Y|X = x \sim N(\alpha + \beta x, \gamma^2)$. Also we know

joint density of (X, Y) = marginal density of $X \times$ conditional density of $Y|X$.

Thus the likelihood of θ based on the n paired observations on (X, Y) and $N - n$ additional observations on X alone, will be

$$L(\theta) = \text{Const}\frac{1}{(\sigma_1^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}(x_i - \mu_1)^2\right\} \times \frac{1}{(\gamma^2)^{n/2}} \exp\left\{-\frac{1}{2\gamma^2}\sum_{j=1}^{n}(y_j - \alpha - \beta x_j)^2\right\}$$

Let $\ell = \ln L(\theta)$. Then the likelihood equations w.r.t. $\mu_1, \sigma_1^2, \alpha, \beta$ and γ^2 will be

$$\frac{\partial \ell}{\partial \mu_1} = 0 \implies \hat{\mu}_1 = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \;\ldots\ldots\; (1)$$

$$\frac{\partial \ell}{\partial \sigma_1^2} = 0 \implies \hat{\sigma}_1^2 = s_1^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \;\ldots\ldots\; (2)$$

$$\frac{\partial \ell}{\partial \alpha} = 0 \implies \hat{\alpha} + \hat{\beta}\overline{x}_n = \overline{y} \;\ldots\ldots\; (3)$$

$$\frac{\partial \ell}{\partial \beta} = 0 \implies \hat{\alpha}\overline{x}_n + \hat{\beta}\frac{1}{n}\sum_{j=1}^{n} x_j^2 = \frac{1}{n}\sum_{j=1}^{n} x_j y_j \;\ldots\ldots\; (4)$$

$$\frac{\partial \ell}{\partial \gamma^2} = 0 \implies \hat{\gamma}^2 = \frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{\alpha} - \hat{\beta}x_j)^2 \;\ldots\ldots\; (5)$$

Equations (1) and (2) directly provide the MLE of μ_1 and σ_1^2. Solutions of (3) and (4) can be expressed as

$$\hat{\beta} = \frac{s_{12n}}{s_{1n}^2} \text{ and } \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}_n$$

where $s_{12n} = \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \overline{x}_n \overline{y}$ and $s_{1n}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)^2$. Also on simplification, (5) gives

$$\hat{\gamma}^2 = s_2^2 - \hat{\beta}^2 s_{1n}^2 \text{ where } s_2^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2.$$

Finally using the relations as in $(*)$ we get $\hat{\mu}_2 = \hat{\alpha} + \hat{\beta}\hat{\mu}_1$, $\hat{\sigma}_2^2 = \hat{\gamma}^2 + \hat{\beta}^2\hat{\sigma}_1^2$ and $\hat{\rho} = \hat{\beta}\frac{\hat{\sigma}_1}{\hat{\sigma}_2}$ as the MLEs of μ_2, σ_2^2 and ρ.

Application of MLE under BVN distribution using R

Let (X, Y) have a BVN distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$. Suppose that 10 observations are made on (X, Y), and additional 5 observations on X only; i.e. 5 observations on Y are missing. These are as follows:

x :	2.95	2.52	1.99	1.40	0.11	1.88	1.46	1.69	2.44	2.60
y :	4.03	2.25	2.01	1.73	1.59	3.60	2.07	2.82	1.12	3.19
x :	1.88	2.48	1.77	3.34	2.75					

Find the MLEs of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$.

Computation using R :

R Code and Output :

```
> x=c(2.95,2.52,1.99,1.4,0.11,1.88,1.46,1.69,2.44,
 2.60,1.88,2.48,1.77,3.34,2.75)
> y=c(4.03,2.25,2.01,1.73,1.59,3.6,2.07,2.82,1.12,
3.19,NA,NA,NA,NA,NA)
> mat=cbind(x,y)# sample observations
> n=10# number of paired observations (X,Y)
> N=15# total number of observations
> sum1=0
> for(i in 1:N)
+ sum1=sum1+x[i]
> MLE_mu1=sum1/N# MLE of mu1
> sum2=0
> for(i in 1:N)
+ sum2=sum2+((x[i]-MLE_mu1)^2)
> MLE_sigma1_2=sum2/N# MLE of sigma1_2
> xn_bar=mean(x[1:n])
> y_bar=mean(na.omit(y))
```

Multivariate Multiparameter MLE

Computation using R :

R Code and Output (continued) :

```
> sum3=0
> for(i in 1:n)
+ sum3=sum3+(x[i]*y[i])
> s12n=(sum3/n)-(xn_bar*y_bar)
> sum4=0
> for(i in 1:n)
+ sum4=sum4+((x[i]-xn_bar)^2)
> s1n_2=sum4/n
> beta=0
> beta=s12n/s1n_2# calculation of beta_hat
> alpha=y_bar-(beta*xn_bar)# calculation of alpha_hat
> sum5=0
> for(i in 1:n)
+ sum5=sum5+((y[i]-y_bar)^2)
> s2_2=sum5/n
> gamma_hat_2=s2_2-((beta^2)*s1n_2)
> MLE_mu2=alpha+(beta*MLE_mu1)# MLE of mu2
> MLE_sigma2_2=gamma_hat_2+((beta^2)*MLE_sigma1_2)# MLE of sigma2_2
> MLE_rho=beta*(sqrt(MLE_sigma1_2)/sqrt(MLE_sigma2_2))# MLE of rho
```

Computation using R :

R Code and Output (continued) :

```
> MLE_mu1
[1] 2.084
> MLE_mu2
[1] 2.539358
> MLE_sigma1_2
[1] 0.5740507
> MLE_sigma2_2
[1] 0.7839372
> MLE_rho
[1] 0.4675964
```

Ex. (*Multivariate multiparameter case*) Let

$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} MN(m, p_1, p_2, \ldots, p_k)$ $0 \leq p_k \leq 1$, $\sum_{i=1}^{k} p_i \leq 1$, a k-variate multinomial distribution.

Assuming m, the multinomial index, as known, the likelihood function of $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ is

$$L(\mathbf{p}) = \text{ Constant } p_1^{\sum_{\alpha=1}^{n} x_{1\alpha}} \ldots . p_k^{\sum_{\alpha=1}^{n} x_{k\alpha}} \left(1 - \sum_{i=1}^{k} p_i\right)^{mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{k} x_{i\alpha}}$$

e-Pathshala पाठशाला MHRD Govt. of India

$\Longrightarrow$ the loglikelihood of $\mathbf{p}$ is

$$\begin{aligned}\ell(\mathbf{p}) &= \text{Constant} + \sum_{\alpha=1}^{n} x_{1\alpha} \log p_1 + \ldots + \sum_{\alpha=1}^{n} x_{k\alpha} \log p_k \\ &+ (mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{k} x_{i\alpha}) \log(1 - \sum_{i=1}^{k} p_i)\end{aligned}$$

Or

$$\frac{\partial}{\partial p_i}\ell(\mathbf{p}) = \frac{\sum_{\alpha=1}^{n} x_{i\alpha}}{p_i} - \frac{mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{k} x_{i\alpha}}{1 - \sum_{i=1}^{k} p_i} \ \forall\ i$$

Multivariate Multiparameter MLE

Hence the likelihood equation : $\frac{\partial}{\partial p_i}\ell(\mathbf{p}) = 0 \ \forall \ i \ \ldots \ \ldots \ (**)$

$$\implies 1 - \sum_{i=1}^{k} p_i = \frac{mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{k} x_{i\alpha}}{mn}$$ (*on simplification by addendo-dividendo method*)

Substituting this form in likelihood equations $(**)$ we get

$$p_i = \frac{\sum_{\alpha=1}^{n}\sum_{i=1}^{k} x_{i\alpha}}{mn} = \frac{\overline{x}_{i.}}{m} = \hat{p}_i \text{ say } \forall \ i$$

where $\overline{x}_{i.} = \frac{1}{n}\sum_{\alpha=1}^{n} x_{i\alpha}$ = sample mean of the ith component variable.

Again, on simplification

$$\frac{\partial^2}{\partial p_i^2}\ell(\mathbf{p})|_{p_i=\hat{p}_i} = -nm^2\left\{\frac{1}{\overline{x}_{i.}} + \frac{1}{m - \sum\limits_{\alpha=1}^{n}\overline{x}_{i.}}\right\} < 0 \ \forall \ i$$

And

$$\frac{\partial^2}{\partial p_i \partial p_j}\ell(\mathbf{p})|_{p_i=\hat{p}_i} = -\frac{nm^2}{m - \sum\limits_{\alpha=1}^{n}\overline{x}_{i.}} < 0 \ \forall \ i, j \ ; \ i \neq j.$$

Now the Hessian matrix is

$$H = \left(\left(\frac{\partial^2}{\partial p_i \partial p_j} l(\mathbf{p})\right)\right) = -\begin{pmatrix} a_1 + b & b & \dots & b \\ b & a_2 + b & \dots & b \\ \vdots & \vdots & \vdots & \vdots \\ b & b & \dots & a_k + b \end{pmatrix}$$

where $a_i = \frac{nm^2}{\overline{x}_{i.}} > 0$ and $b = \frac{\frac{nm^2}{n}}{m - \sum\limits_{\alpha=1} \overline{x}_{i.}} > 0.$

Multivariate Multiparameter MLE

Claim: H is negative definite.

Verification. Define

$$H^* = -H = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & a_k \end{pmatrix} + b\,\mathbf{1}\mathbf{1}'$$

Then

$$\mathbf{y}'H^*\mathbf{y} = \sum_{i=1}^{n} a_i y_i^2 + b(\sum_{i=1}^{n} y_i)^2 \geq 0 \text{ with } = \text{ iff } \mathbf{y} = \mathbf{0}.$$

$\implies$ H^* is positive definite $\implies$ H is negative definite.

$\implies$ $\{\hat{p}_i \; i = 1, \dots, k\}$ maximises the likelihood function and hence is the MLE of $\mathbf{p}$.

Multivariate Multiparameter MLE

e-Pathshala पाठशाला MHRD Govt. of India

TRY YOURSELF!

M5.1. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample of size n drawn from a continuous $p-$variate distribution with common joint density

$$
\begin{aligned}
f(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}\prod_{i=1}^{p} x_i} \exp\left[-\frac{1}{2}(\ln \mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\ln \mathbf{x} - \boldsymbol{\mu})\right] \;,\; \mathbf{x} > \mathbf{0},\; \Sigma > \\
&= 0 \quad \text{Otherwise}
\end{aligned}
$$

where
$\mathbf{x} = (x_1, \ldots, x_p) > \mathbf{0} \Rightarrow x_i > 0 \;\forall\; i$, $\ln \mathbf{x} = (\ln x_1, \ldots, \ln x_p)$, $\Sigma > O$ means Σ is a $p \times p$ symmetric p.d. matrix and $\boldsymbol{\mu} \in \mathcal{R}^p$.

Find the MLE of $(\boldsymbol{\mu}, \Sigma)$.

M5.2. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample of size n drawn from a multinomial population MN $(m; p_1, p_2, p_3, p_2, p_1)$ such that $2p_1 + 2p_2 + p_3 \leq 1$. Find the ML estimator of $(p_1, p_2, p_3)'$.

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 5 ...

M5.1. The given distribution is called Multivariate ($p-$ variate) lognormal distribution with parameters $(\boldsymbol{\mu}, \Sigma)$. This is usually denoted by $LN_p(\boldsymbol{\mu}, \Sigma)$.

As in the case of univariate distribution, in multivariate case also there is a relation between multivariate lognormal and multivariate normal distributions. This is as follows

Let $\mathbf{X} = (X_1, \ldots, X_p)' \sim \; LN_p(\boldsymbol{\mu}, \Sigma)$.

Then $\ln \mathbf{X} = (\ln X_1, \ldots, \ln X_p)' \sim \; N_p(\boldsymbol{\mu}, \Sigma)$.

Proof of this result is very much straightforward.

Multivariate Multiparameter MLE

Make a transformation from $(X_1, \ldots, X_p) \longrightarrow (Y_1, \ldots, Y_p)$ where $Y_i = \ln X_i$.

Now using Jacobian of transformation or otherwise, get the joint p.d.f. of $\mathbf{Y} = (Y_1, \ldots, Y_p)$. You will find that

$$\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$$

Following this result we can say that in problem **M5.1.**

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \Sigma)$$

where $\mathbf{Y}_\alpha = \ln \mathbf{X}_\alpha \ , \ \alpha = 1, \ldots, n.$

Now following the ML estimation under multivariate normal distribution, we get the MLE of $(\boldsymbol{\mu}, \Sigma)$ is $(\overline{\mathbf{Y}}, S_Y)$ where $\overline{\mathbf{Y}}$ = sample mean vector and S_Y = sample variance-covariance matrix on $\mathbf{Y}$.

M5. 2. Here $\mathbf{X}_\alpha = (X_{1\alpha}, X_{2\alpha}, X_{3\alpha}, X_{4\alpha}, X_{5\alpha})'$, $\alpha = 1, \ldots, n$.

The likelihood of $\{p_1, p_2, p_3\}$ is

$$L(p_1, p_2, p_3) = \text{Const } p_1^{\sum_{\alpha=1}^{n}(x_{1\alpha}+x_{5\alpha})} p_2^{\sum_{\alpha=1}^{n}(x_{2\alpha}+x_{4\alpha})} p_3^{\sum_{\alpha=1}^{n} x_{3\alpha}}$$

$$+ \ (1 - 2p_1 - 2p_2 - p_3)^{(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5} x_{i\alpha})}$$

Multivariate Multiparameter MLE

$$\frac{\partial}{\partial p_1}\ell(\mathbf{p}) = \frac{\sum_{\alpha=1}^{n}(x_{1\alpha}+x_{5\alpha})}{p_1} - \frac{2(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5}x_{i\alpha})}{1-2p_1-2p_2-p_3}$$

$$\frac{\partial}{\partial p_2}\ell(\mathbf{p}) = \frac{\sum_{\alpha=1}^{n}(x_{2\alpha}+x_{4\alpha})}{p_2} - \frac{2(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5}x_{i\alpha})}{1-2p_1-2p_2-p_3}$$

and

$$\frac{\partial}{\partial p_3}\ell(\mathbf{p}) = \frac{\sum_{\alpha=1}^{n} x_{3\alpha}}{p_3} - \frac{(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5} x_{i\alpha})}{1 - 2p_1 - 2p_2 - p_3}$$

Hence $\frac{\partial}{\partial p_i}\ell(\mathbf{p}) = 0 \ \forall \ i = 1, 2, 3 \implies$

$$\frac{2p_1}{1 - 2p_1 - 2p_2 - p_3} = \frac{\sum_{\alpha=1}^{n}(x_{1\alpha} + x_{5\alpha})}{(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5} x_{i\alpha})}$$

$$\frac{2p_2}{1-2p_1-2p_2-p_3} = \frac{\sum_{\alpha=1}^{n}(x_{2\alpha}+x_{4\alpha})}{(mn-\sum_{\alpha=1}^{n}\sum_{i=1}^{5}x_{i\alpha})}$$

And

$$\frac{p_3}{1-2p_1-2p_2-p_3} = \frac{\sum_{\alpha=1}^{n}x_{3\alpha}}{(mn-\sum_{\alpha=1}^{n}\sum_{i=1}^{5}x_{i\alpha})}$$

Denote these three equations by $(*)$.

Multivariate Multiparameter MLE

Now $(*) \implies$

$$1 - 2p_1 - 2p_2 - p_3 = \frac{(mn - \sum_{\alpha=1}^{n}\sum_{i=1}^{5} x_{i\alpha})}{mn}$$

Substituting this in $(*) \implies$

$$p_1 = \frac{1}{2mn}\sum_{\alpha=1}^{n}(x_{1\alpha} + x_{5\alpha}) = \frac{1}{2m}(\overline{X}_{1.} + \overline{X}_{5.}) = \hat{p}_1$$

$$p_2 = \frac{1}{2mn}\sum_{\alpha=1}^{n}(x_{2\alpha} + x_{4\alpha}) = \frac{1}{2m}(\overline{X}_{2.} + \overline{X}_{4.}) = \hat{p}_2,$$

And

$$p_3 = \frac{1}{mn}\sum_{\alpha=1}^{n} x_{3\alpha} = \frac{\overline{X}_{3.}}{m} = \hat{p}_3.$$

Now straightway find the Hessian matrix and it is not difficult to show that the Hessian matrix is n.d. at $(p_1, p_2, p_3) = (\hat{p}_1, \hat{p}_2, \hat{p}_3)$

$\implies$ the likelihood is maximum at $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$.

$\implies$ $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$: MLE of (p_1, p_2, p_3).

One Parameter Exponential Family and MLE

Module 6

Saurav De

Department of Statistics
Presidency University

One Parameter Exponential Family (OPEF)

Suppose, based on a random sample of size n the joint pmf (or pdf) can be expressed as $p_\theta(\mathbf{x}) = \exp\left[Q(\theta)T(\mathbf{x}) + c(\theta) + D(\mathbf{x})\right]$, θ, real valued.

Assumptions:

1. The first two derivatives of $Q(\theta)$ and $c(\theta)$ exist and are continuous.
2. $I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log L\right]^2$ exists and is positive.

where L is the likelihood function of θ based on the random sample.

Then $P = \{p_\theta(\mathbf{x}) : \theta \in \Omega\}$ is called OPEF.

One Parameter Exponential Family and MLE

Check that $E_\theta T(\mathbf{X}) = -\frac{c'(\theta)}{Q'(\theta)}$

Hint. Here $p_\theta(\mathbf{x}) = \exp\left[Q(\theta)T(\mathbf{x}) + c(\theta) + D(\mathbf{x})\right]$

$$
\begin{aligned}
E_\theta(Q'(\theta)T(\mathbf{X}) + c'(\theta)) &= \int \left\{Q'(\theta)T(\mathbf{x}) + c'(\theta)\right\} p_\theta(\mathbf{x})d(\mathbf{x}) \\
&= \int \frac{\partial}{\partial\theta}\left\{\exp\left[Q(\theta)T(\mathbf{x}) + c(\theta) + D(\mathbf{x})\right]\right\} d(\mathbf{x}) \\
&= \frac{\partial}{\partial\theta}\int p_\theta(\mathbf{x})d(\mathbf{x}) \text{ as } \int \text{ and } \frac{\partial}{\partial\theta} \text{ interchangeable} \\
&= \frac{\partial}{\partial\theta}(1) = 0
\end{aligned}
$$

$$\implies E_\theta T(\mathbf{X}) = -\frac{c'(\theta)}{Q'(\theta)}$$

One Parameter Exponential Family and MLE

$$V_\theta T(\mathbf{X}) = \left\{ \frac{Q''(\theta)c'(\theta)}{Q'(\theta)} - c''(\theta) \right\} \frac{1}{(Q'(\theta)^2)}.$$

Hint. $\frac{\partial^2}{\partial\theta^2} \int p_\theta(\mathbf{x})d(\mathbf{x}) = \frac{\partial^2}{\partial\theta^2}(1) = 0$

Or $\int \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} p_\theta(\mathbf{x})d(\mathbf{x}) = 0$ (*as* $\int$ and $\frac{\partial^2}{\partial\theta^2}$ *are interchangeable*)

$$\implies \int \frac{\partial}{\partial\theta} \left\{ Q'(\theta)T(\mathbf{x}) + c'(\theta) \right\} p_\theta(\mathbf{x})d(\mathbf{x}) = 0$$

$$\Leftrightarrow \int (Q''(\theta)T(\mathbf{x}) + c''(\theta))p_\theta(\mathbf{x})d(\mathbf{x}) + (Q'(\theta)T(\mathbf{x}) + c'(\theta))^2 p_\theta(\mathbf{x})d(\mathbf{x}) = 0$$

One Parameter Exponential Family and MLE

Or

$$E_\theta[Q''(\theta)T(\mathbf{X}) + c''(\theta)] + E_\theta[Q'(\theta)T(\mathbf{x}) + c'(\theta)]^2 = 0$$

Or

$$Q''(\theta)E_\theta\left(T(\mathbf{X})\right) + c''(\theta) + (Q'(\theta))^2 E_\theta\left[T(\mathbf{X}) - \left(-\frac{c'(\theta)}{Q'(\theta)}\right)\right]^2 = 0$$

Note that 2nd term in LHS $= (Q'(\theta))^2 V_\theta\left(T(\mathbf{X})\right)$.

Hence get $V_\theta\left(T(\mathbf{X})\right)$.

One Parameter Exponential Family and MLE

The likelihood equation for the probability model under OPEF is

$$\frac{\partial}{\partial\theta}\ln L = c'(\theta) + Q'(\theta)T(\mathbf{x}) = 0$$

Or $T(\mathbf{x}) = -\frac{c'(\theta)}{Q'(\theta)}$.

Note. In particular for $n = 1$ the pmf or pdf $f_\theta(x) \in$ OPEF if

$$f_\theta(x) = \exp\left[Q(\theta)T^*(x) + \psi(\theta) + h(x)\right]\ ,\ \theta \in \Omega(\subseteq \mathcal{R})$$

satisfying the abovementioned assumptions. Then

$$E_\theta T^*(X) = -\frac{\psi'(\theta)}{Q'(\theta)}$$

One Parameter Exponential Family and MLE

With this form of common density the joint pdf of n independent random sample observations will be naturally

$$p_\theta(\mathbf{x}) = \exp\left[Q(\theta)\sum_{i=1}^{n} T^*(x_i) + n\psi(\theta) + \sum_{i=1}^{n} h(x_i)\right]$$

Without loss of generality, it can be expressed as

$$p_\theta(\mathbf{x}) = \exp\left[Q(\theta)T(\mathbf{x}) + c(\theta) + D(\mathbf{x})\right]$$

with $T(\mathbf{x}) = \sum_{i=1}^{n} T^*(x_i)$, $c(\theta) = n\psi(\theta)$ and $D(\mathbf{x}) = \sum_{i=1}^{n} h(x_i)$

Clearly $p_\theta(\mathbf{x}) \in$ OPEF.

One Parameter Exponential Family and MLE

Result. The method of moments and the method of maximum likelihood agree each other for OPEF distributions.

Let the common pdf(or pmf): $f_\theta(x) = \exp\left[Q(\theta)T^*(x) + \psi(\theta) + h(x)\right] \in$ OPEF. Then, based on a random sample of size n the loglikelihood function : $l_{\mathbf{x}}(\theta) = Q(\theta)\sum_{i=1}^{n} T^*(x_i) + n\psi(\theta) + D(\mathbf{x})$.

Now

$$\frac{d}{d\theta} l_{\mathbf{x}}(\theta) = 0 \;\Rightarrow\; \frac{\psi'(\theta)}{Q'(\theta)} = -\frac{\sum_{i=1}^{n} T^*(x_i)}{n} \;\Rightarrow\; E_\theta T^*(X) = \frac{\sum_{i=1}^{n} T^*(X_i)}{n} \quad \ldots(*)$$

But $(*)$ is the moment equation with respect to the random variable $T^*(X)$. Hence proved.

One Parameter Exponential Family and MLE

Result. For any distribution in OPEF

(i) any solution of the likelihood equation provides a maximum of likelihood function.

(ii) a solution of likelihood equation, if exists, is unique.

(i) and (ii) $\implies$ the solution of the likelihood equation is unique MLE.

Proof. (i) Let $\tilde{\theta}$ be a solution of the likelihood equation. Then

$$c'(\tilde{\theta}) + Q'(\tilde{\theta})T(\mathbf{x}) = 0 \implies -\frac{c'(\tilde{\theta})}{Q'(\tilde{\theta})} = T(\mathbf{x})$$

One Parameter Exponential Family and MLE

Now $-I(\theta) = E\left(\frac{\partial^2}{\partial\theta^2} l_{\mathbf{x}}(\theta)\right) = c''(\theta) + Q''(\theta)E(T(\mathbf{X})) < 0 \; \forall\, \theta$ as $I(\theta) > 0$.

$$\implies c''(\theta) - Q''(\theta)\frac{c'(\theta)}{Q'(\theta)} < 0 \; \forall\, \theta$$

But $\frac{\partial^2}{\partial\theta^2} l_{\mathbf{x}}(\theta)|_{\theta=\tilde{\theta}} = c''(\tilde{\theta}) - Q''(\tilde{\theta})\frac{c'(\tilde{\theta})}{Q'(\tilde{\theta})}$.

$$\implies \frac{\partial^2}{\partial\theta^2} l_{\mathbf{x}}(\theta)|_{\theta=\tilde{\theta}} < 0$$

So $\tilde{\theta}$ maximises the likelihood function of θ.

(ii) If possible suppose $\exists$ another solution $\tilde{\tilde{\theta}}$ of the likelihood equation.

$(i) \implies \tilde{\tilde{\theta}}$ also maximises the likelihood function, like $\tilde{\theta}$.

$\implies \exists$ another solution of the likelihood equation in between $\tilde{\theta}$ and $\tilde{\tilde{\theta}}$ which minimises the likelihood function.

$\implies$ contradiction to (i). Hence our supposition is wrong. In other words the solution of the likelihood equation, if exists, is only one i.e. unique.

One Parameter Exponential Family and MLE

Notes.

1. • We know for OPEF distributions, $T(\mathbf{X})$ is the complete sufficient statistic.

   • Also here the MLE is the unique solution of $T(\mathbf{x}) = -\frac{c'(\theta)}{Q'(\theta)}$.

   $\Longrightarrow$ the MLE and the complete sufficient statistic $T(\mathbf{X})$ are in $1:1$ relation.
   But we know by Rao-Blackwell-Lehmann-Scheffe Theorem that MVUE is a function of complete sufficient statistic.

   $\Longrightarrow$ under OPEF the MVUE can be obtained from the MLE just by bias correction.

One Parameter Exponential Family and MLE

2. We get $\frac{\partial^2}{\partial\theta^2} \log L|_{\tilde{\theta}} = -I(\tilde{\theta})$ under OPEF

$$\implies I(\theta) = -\frac{\partial^2}{\partial\theta^2} \log L|_{\tilde{\theta}=\theta}$$

This helps evaluate $I(\theta)$ avoiding mathematical expectation. In fact the equivalence of the moment equation and the likelihood equation under OPEF is responsible for this.

We illustrate this interesting matter through the next example.

One Parameter Exponential Family and MLE

Ex. $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \theta)$

$$\log L = \text{ const } - \frac{n}{2} \log \theta - \frac{1}{2\theta} \sum x_i^2$$

$$\frac{\partial}{\partial \theta} \log L = 0 \implies -\frac{n}{2\theta} + \frac{\sum x_i^2}{2\theta^2} = 0 \implies \tilde{\theta} = \frac{1}{n} \sum x_i^2$$

where $\tilde{\theta}$ is the unique MLE of θ.

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \log L &= \frac{n}{2\theta^2} - \frac{\sum x_i^2}{\theta^3} = \frac{n}{\theta^3} \left[\frac{\theta}{2} - \frac{\sum x_i^2}{n} \right] \\ &= \frac{n}{\theta^3} \left[\frac{\theta}{2} - \tilde{\theta} \right] \end{aligned}$$

$$\implies I(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L|_{\tilde{\theta} = \theta} = -\frac{n}{\theta^3} \left[\frac{\theta}{2} - \theta \right] = \frac{n}{2\theta^2}.$$

One Parameter Exponential Family and MLE

3. Let $\exists$ an unbiased estimator $T(\mathbf{X})$ of $g(\theta)$ with variance attaining **CRLB**. Then $p_\theta(\mathbf{x})$ is of the exponential form and

$$\frac{\partial}{\partial\theta}\log L = k(\theta)(T(\mathbf{X}) - g(\theta))$$

$\implies$ the unique MLE of $g(\theta)$ is $T(\mathbf{X})$.

4. Our discussion can straightway be extended to Multi-parameter Exponential Family (MPEF). The results are very much similar to those obtained under OPEF. For detailed study consult **A First Course on Parametric Inference** by *B. K. Kale*.

Power Series Distribution Family

Let $X \sim$ discrete probability distribution with p.m.f. f_θ of the form

$$
\begin{aligned}
f_\theta(x) &= \frac{a_x \theta^x}{g(\theta)} \text{ if } x = 0, 1, \ldots \\
&= 0 \text{ Otherwise}
\end{aligned}
$$

where $\theta > 0$, a_x is a positive real-valued function of x and $g(\theta)$ is a positive real-valued function of θ. Any discrete probability distribution of this form is called Power Series Distribution. The corresponding family is called Power Series Distribution family.

One Parameter Exponential Family and MLE

As $f_\theta(x)$ is a p.m.f. at the point x,
$\sum_{x\geq 0} f_\theta(x) = 1 \implies g(\theta) = \sum_{x\geq 0} a_x\theta^x$, $\theta > 0$. Now

$$\begin{aligned}
E(X) &= \sum_{x\geq 1} x\frac{a_x\theta^x}{g(\theta)} \\
&= \theta\sum_{x\geq 1} x\frac{a_x\theta^{x-1}}{g(\theta)} \\
&= \frac{\theta}{g(\theta)}\sum_{x\geq 0}\frac{\partial}{\partial\theta}\left\{a_x\theta^x\right\} \\
&= \frac{\theta}{g(\theta)}\frac{\partial}{\partial\theta}\sum_{x\geq 0} a_x\theta^x
\end{aligned}$$

One Parameter Exponential Family and MLE

$$= \frac{\theta}{g(\theta)} \frac{\partial}{\partial\theta} g(\theta)$$

$$= \theta \frac{\partial}{\partial\theta} \ln g(\theta)$$

If $X_1, \ldots, X_n$ be n random observations on X using method of moments we get the moment equation

$$E(X) = \overline{x}\,, \text{ sample mean}$$

Or

$$\theta \frac{\partial}{\partial\theta} \ln g(\theta) = \overline{x} \quad \ldots \quad \ldots \quad (*)$$

One Parameter Exponential Family and MLE

On the other hand the likelihood function of θ is

$$L(\theta) = \frac{\left(\prod_{i=1}^{n} a_{x_i}\right) \theta^{\sum_{i=1}^{n} x_i}}{(g(\theta))^n}$$

This implies the loglikelihood function of θ is

$$\ell(\theta) = \text{ const } + \sum_{i=1}^{n} x_i \ln\theta - n \ln g(\theta)$$

Now

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - n\frac{1}{g(\theta)}\frac{\partial}{\partial\theta}g(\theta)$$

One Parameter Exponential Family and MLE

Or

$$\frac{\partial}{\partial\theta}\ell(\theta) = n\frac{\overline{x}}{\theta} - n\frac{\partial}{\partial\theta}\ln g(\theta)$$

Hence the likelihood equation is

$$\frac{\partial}{\partial\theta}\ell(\theta) = 0 \iff \frac{\overline{x}}{\theta} = \frac{\partial}{\partial\theta}\ln g(\theta) \quad \ldots \quad \ldots \quad (**)$$

$(*)$ and $(**)$ are same, $\implies$ the method of moments and the method of maximum likelihood coincide for power series distribution.

One Parameter Exponential Family and MLE

Note. Depending on the choices of a_x , θ and $g(\theta)$ functions, we get a family of power series distributions.

One of the well-known members of this family is Bernoulli distribution [for the choice $a_x = 1$ for $x = 0, 1$, $\theta = \frac{p}{1-p}$ and $g(\theta) = (1-p)^{-1} = (1+\theta)$]. For this distribution

$$\ln g(\theta) = \ln(1+\theta) \implies \frac{\partial}{\partial\theta} \ln g(\theta) = (1+\theta)^{-1} = (1-p)$$

One Parameter Exponential Family and MLE

Hence the likelihood equation $\theta \frac{\partial}{\partial\theta} \ln g(\theta) = \overline{x}$ becomes

$$\frac{p}{1-p}(1-p) = \overline{x} \iff p = \overline{x}$$

This is also the solution of the likelihood as well as moment equation. In fact $\hat{p} = \overline{X}$ is the MLE as well as the MME (Method of Moment Estimator) of p under Bernoulli(p) distribution.

Similarly Poisson(λ) distribution is another member with the choice $a_x = \frac{1}{x!}$, $\theta = \lambda$ and $g(\theta) = \exp[\theta]$.

Here also we can verify in the similar way that $\overline{X}$ is the MLE as well as the MME of λ.

One Parameter Exponential Family and MLE

Polynomial Type Exponential Distribution and MLE

A random variable X has polynomial type exponential distribution if its density is defined as

$$f_\theta(x) = \exp[-\sum_{i=0}^{m} \theta_i x^i] \; ; \; x \; > 0$$

where the exponent is a polynomial in x of degree at the most m and any one parameter say θ_0 is a function of the remaining parameters.

As

$$\int \exp[-\theta_0 - \theta_1 x - \theta_2 x^2 - \ldots - \theta_m x^m] dx = 1 \implies \exp[\theta_0] = \int \exp[-\sum_{i=1}^{m} \theta_i x^i]$$

One Parameter Exponential Family and MLE

Then the rth population raw moment is

$$\begin{aligned}
\mu'_r = E(X^r) &= \int x^r \exp[-\sum_{i=0}^{m}\theta_i x^i]dx\ , \ \ r = 1, 2, \ldots \\
&= \exp[-\theta_0] \int \frac{\partial}{\partial\theta_r} \exp[-\sum_{i=1}^{m}\theta_i x^i]dx \\
&= \left(\int \exp[-\sum_{i=1}^{m}\theta_i x^i]dx\right)^{-1} \frac{\partial}{\partial\theta_r} \int \exp[-\sum_{i=1}^{m}\theta_i x^i]dx \\
&= \frac{\partial}{\partial\theta_r} \ln \int \exp[-\sum_{i=1}^{m}\theta_i x^i]dx = \frac{\partial}{\partial\theta_r} \exp[\theta_0]
\end{aligned}$$

One Parameter Exponential Family and MLE

Let $X_1, \ldots, X_n$ be drawn from above distribution. Then using Method of Moments m moment equations are

$$\mu'_r = m'_r \left(= \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha^r \right) \;,\; r = 1, 2, \ldots, m$$

That is

$$\frac{\partial}{\partial \theta_r} \ln \int \exp[-\sum_{i=1}^{m} \theta_i x^i] dx = m'_r \;,\; r = 1, 2, \ldots, m \quad \ldots \quad \ldots \quad (*)$$

On the other hand the likelihood function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is

$$L(\boldsymbol{\theta}) = \exp\{-n\,\theta_0\} \exp\left\{ -\sum_{i=1}^{m} \theta_i \sum_{\alpha=1}^{n} x_\alpha^i \right\}$$

Or

$$L(\boldsymbol{\theta}) = \exp\{-n\,\theta_0\} \exp\left\{ -n \sum_{i=1}^{m} \theta_i m'_i \right\}$$

One Parameter Exponential Family and MLE

Now

$$\frac{\partial}{\partial\theta_r}L(\boldsymbol{\theta}) = -n\frac{\partial}{\partial\theta_r}(\theta_0)\exp\{-n\,\theta_0\}\exp\left\{-n\sum_{i=1}^{m}\theta_i m_i'\right\} \\ -n\,m_r'\exp\{-n\,\theta_0\}\exp\left\{-n\sum_{i=1}^{m}\theta_i m_i'\right\}$$

$\frac{\partial}{\partial\theta_r}L(\boldsymbol{\theta}) = n\left(-\frac{\partial}{\partial\theta_r}(\theta_0) - m_r'\right)\exp\{-n\,\theta_0\}\exp\left\{-n\sum_{i=1}^{m}\theta_i m_i'\right\}\ r = 1, 2, \ldots, m$

Or $\frac{\partial}{\partial\theta_r}L(\boldsymbol{\theta}) =$
$n\left(\frac{\partial}{\partial\theta_r}\ln\int\exp[-\sum_{i=1}^{m}\theta_i x^i]dx - m_r'\right)\exp\{-n\,\theta_0\}\exp\left\{-n\sum_{i=1}^{m}\theta_i m_i'\right\}\ r = 1, 2, \ldots, m$

One Parameter Exponential Family and MLE

Hence the likelihood equations are

$$\frac{\partial}{\partial\theta_r}L(\boldsymbol{\theta}) = 0 \;\; r = 1, 2, \ldots, m$$

$\Longleftrightarrow$

$$\frac{\partial}{\partial\theta_r}\ln\int \exp[-\sum_{i=1}^{m}\theta_i x^i]dx = m'_r\ , \;\; r = 1, 2, \ldots, m \;\; \ldots \;\; \ldots \;\; (**)$$

$$[\text{as } \exp\{-n\,\theta_0\}\exp\left\{-n\sum_{i=1}^{m}\theta_i m'_i\right\} > 0.]$$

Since $(*)$ and $(**)$ are ientical, again for this distribution also the method of moments and the method of maximum likelihood agree to each other.

TRY YOURSELF !

M6. 1. let $X_1, \ldots, X_n$ independently follow negative binomial distribution with common pmf

$$f(x) = {}^{r+x-1}C_x (1-\theta)^r \theta^x \ , \ x = 0, 1, \ldots ; 0 \leq \theta \leq 1.$$

Show that the distribution $\in$ OPEF and hence find the MLE of θ.

Also get the Fisher's Information for θ.

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 6 . . .

M6. 1. Choosing $a_x = ^{r+x-1} C_x$ and $g(\theta) = (1-\theta)^{-r}$ pmf of the given negative binomial distribution becomes $f(x) = \frac{a_x \theta^x}{g(\theta)}$ which is a power series distribution and hence $\in$ OPEF.

Also here $\theta \frac{\partial}{\partial \theta} g(\theta) = \frac{r\theta}{1-\theta}$.

So from the equation $\theta \frac{\partial}{\partial \theta} g(\theta) = \overline{x}$ (*which is moment as well as likelihood equation*) we get $\theta = \frac{\overline{x}}{r+\overline{x}}$.

$\Longrightarrow$ MLE of θ : $\frac{\overline{X}}{r+\overline{X}} = \tilde{\theta}$ (say)

One Parameter Exponential Family and MLE

Check that here the loglikelihood of θ is

$$\ell(\theta) = \text{Const} + nr\ln(1-\theta) + \sum x_i \ln\theta$$

Then it is easy to verify that

$$\begin{aligned}
-\frac{\partial^2}{\partial\theta^2}\ell(\theta) &= \frac{nr}{(1-\theta)^2} + \frac{n\overline{x}}{\theta^2} \\
&= \frac{nr}{\theta^2}\left[\frac{\theta^2}{(1-\theta)^2} + \frac{\overline{x}}{r}\right] \\
&= \frac{nr}{\theta^2}\left[\frac{\theta^2}{(1-\theta)^2} + \frac{\tilde{\theta}}{1-\tilde{\theta}}\right]
\end{aligned}$$

So Fisher's Information $I(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta)|_{\tilde{\theta}=\theta} = \frac{nr}{\theta(1-\theta)^2}$ (*on simplification*)

Some Important Theorems on MLE

Module 7

Saurav De

Department of Statistics
Presidency University

Theorem 1.

Assumptions

(i) $X_1, X_2, \ldots, X_n$ are iid $\sim$ p.d.f.$f_\theta(x)$

(ii) The probability distributions $\{P_\theta : \theta \in \Omega\}$ are distinct [i.e. there doesn't exist any $\theta_1 \neq \theta_2$ such that $P_{\theta_1} = P_{\theta_2}$]

(iii) The probability distribution $\{P_\theta : \theta \in \Omega\}$ have common support [for example not applicable for $R(0, \theta)$ distribution because there $P_1 = R(0, 1)$; $P_2 = R(0, 2)$ don't have common support]

Then $P_{\theta_0}\left[L_x(\theta_0) > L_x(\theta)\right] \to 1$ as $n \to \infty$($\theta \neq \theta_0$is fixed).

Some Important Theorems on MLE

Proof. $L_{\mathbf{x}}(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$

So $l_{\mathbf{x}}(\theta) - l_{\mathbf{x}}(\theta_0) = \sum_{i=1}^{n} ln \frac{f_\theta(x_i)}{f_{\theta_0}(x_i)}$.

Now $\frac{1}{n}\sum_{i=1}^{n} ln \frac{f_\theta(x_i)}{f_{\theta_0}(x_i)} \xrightarrow{P} E_{\theta_0} ln \frac{f_\theta(X)}{f_{\theta_0}(X)}$ $(by\ WLLN\ since\ X_1, X_2, \ldots, X_n\ are\ iid)$

$$\leq lnE_{\theta_0} \frac{f_\theta(X)}{f_{\theta_0}(X)} \quad (by\ Jensen's\ inequality)$$

$$= ln1 = 0$$

So $P_{\theta_0}\{L_{\mathbf{x}}(\theta) < L_{\mathbf{x}}(\theta_0)\} \to 1$ as $n \to \infty$. Hence proved.

Theorem 2.

Assumptions: (i) - (iii) [*same as that in* **Theorem 1.**]

(iv) Ω is open.

(v) $l_x(\theta)$ is differentiable everywhere in Ω for almost all x.

Then likelihood equation $\frac{\partial}{\partial\theta} l_x(\theta) = 0$, has one root which is consistent for θ with probability tending to 1.

Proof. Let θ_0 be the true value of θ.

As Ω is open $\exists\ \delta > 0 \ni (\theta_0 - \delta, \theta_0 + \delta) \subset \Omega$.

Define $S_n = \{x : l_x(\theta_0) > l_x(\theta_0 - \delta)\ \&\ l_x(\theta_0) > l_x(\theta_0 + \delta)\}$

$\Longrightarrow$

$$\begin{aligned}
P_{\theta_0}(S_n) \quad &\geq \quad P_{\theta_0}\left[l_x(\theta_0) > l_x(\theta_0 - \delta)\right] + P_{\theta_0}\left[l_x(\theta_0) > l_x(\theta_0 + \delta)\right] - 1 \\
&\qquad \left[\text{using } P(A \cap B) \geq P(A) + P(B) - 1\right] \\
&\longrightarrow \quad 1 \text{ as } n \longrightarrow \infty \,(\text{by Theorem 1}) \ldots\ \ldots\ (\oplus)
\end{aligned}$$

So given x, $\exists \hat{\theta}_n$, a root of the likelihood equation $\ni \; \hat{\theta}_n \in (\theta_0 - \delta, \theta_0 + \delta)$.

So

$$l_x(\theta_0) > l_x(\theta_0 - \delta) \;\&\; l_x(\theta_0) > l_x(\theta_0 + \delta) \implies \hat{\theta}_n \in (\theta_0 - \delta, \theta_0 + \delta)$$

Or

$$P_{\theta_0}(S_n) \leq P_{\theta_0}\left[|\hat{\theta}_n - \theta_0| < \delta\right]$$

So $(\oplus) \implies \hat{\theta}_n \xrightarrow{P} \theta_0$ with probability tending to 1.

Proved

Implication of **Theorem 2** : With probability tending to 1 as $n \to \infty$ $\exists$ a local maxima $\hat{\theta}_n$ which is consistent for θ and hence $l'_x(\hat{\theta}_n) = 0$.

Result: If the likelihood equation has unique root $\hat{\theta}_n$, then with probability tending to 1 as $n \to \infty$ $\hat{\theta}_n$ is the MLE of θ and it is consistent.

Proof. By **Theorem 2** $\hat{\theta}_n$ is a local maxima of $l_x(\theta)$. If possible suppose $\exists$ a $\theta^* \ni l_x(\hat{\theta}_n) < l_x(\theta^*)$.But $\hat{\theta}_n$ is a local maxima of $l_x(\theta)$. Hence $\exists$ a point θ^{**} for which $l_x(\theta)$ is locally minimum.

Hence $\frac{d}{d\theta} l_x(\theta)|_{\theta=\theta^{**}} = 0$.

So $l'_x(\theta) = 0$ has another solution at $\theta = \theta^{**}$, which is a contradiction as $\hat{\theta}_n$ is unique consistent.

$\implies l_x(\hat{\theta}_n) \geq l_x(\theta^*)\ \forall\ \theta^*$. That means $\hat{\theta}_n$ is global minima.

Theorem 3. (*asymptotic normality*)

Asuumtions: (i) - (iv) [*same as that in* **Theorem 2.**]

(v) $l_x(\theta)$ is thrice differentiable with respect to θ and the derivative is continuous.

(vi) $\int f_\theta(x)dx$ can be twice differentiable under the integral sign.

(vii) For every $\theta_0 \in \Omega$, $\exists$ $\delta > 0$ and $M(x) \ni$

$$\left| \frac{\partial^3}{\partial\theta^3} lnf_\theta(x) \right| \leq M(x) \ \forall \ \theta \in (\theta_0 - \delta, \theta_0 + \delta) \quad \forall \ x$$

with $E_{\theta_0} M(x) < \infty$.

(viii) $0 < i(\theta) = E_\theta \left(-\frac{\partial^2}{\partial\theta^2} lnf_\theta(X) \right) = E_\theta \left(\frac{\partial}{\partial\theta} lnf_\theta(X) \right)^2 < \infty$.

Then any consistent solution of the likelihood equation $\hat{\theta}_n$ satisfies

$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)})$.

Proof. From Taylor's expansion of $l'_x(\hat{\theta}_n) = \frac{\partial}{\partial\theta} l_x(\theta)|_{\hat{\theta}_n}$ about θ_0 we get

$$l'_x(\hat{\theta}_n) = l'_x(\theta_0) + (\hat{\theta}_n - \theta_0) l''_x(\theta_0) + \frac{(\hat{\theta}_n - \theta_0)^2}{2} l'''_x(\theta') \ldots \quad \ldots \quad (*)$$

[θ' lies between $\hat{\theta}_n$ and θ_0]

As $\hat{\theta}_n$ is a solution of the likelihood equation $\implies$ $l'_x(\hat{\theta}_n) = 0$.

So $(*)$ $\implies$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} l'_x(\theta_0)}{-\frac{1}{n} l''_x(\theta_0) - \frac{(\hat{\theta}_n - \theta_0)}{2} \frac{1}{n} l'''_x(\theta')}$$
$$= \frac{N}{D_1 + D_2} \text{ say}$$

Now

$$N = \frac{1}{\sqrt{n}} l'_x(\theta_0)$$
$$= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_\theta(x_i)|_{\theta=\theta_0}$$

As X_i's are *iid* so also are $\frac{\partial}{\partial\theta} \ln f_\theta(X_i)|_{\theta=\theta_0}$.

Moreover note that $E_{\theta_0}\left(\frac{\partial}{\partial\theta} \ln f_\theta(X_1)|_{\theta=\theta_0}\right) = 0$ and
$V_{\theta_0} = E_{\theta_0}\left(\frac{\partial}{\partial\theta} \ln f_\theta(X_1)|_{\theta=\theta_0}\right)^2 = i(\theta_0) > 0.$

$\implies$ by Lindeberg-Levy CLT [$Y_1, Y_2, \ldots$ be a sequence of iid random variables with common mean μ and common variance
$\sigma^2 \implies \sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$; $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$]

$$N = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} \ln f_\theta(x_i)|_{\theta=\theta_0} \xrightarrow{\mathcal{D}} N(0, i(\theta_0)) \ldots \quad \ldots \quad (**)$$

Next $D_1 = -\frac{1}{n} l_x''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X_1)|_{\theta_0}\right) \xrightarrow{P} i(\theta_0)$ (*by Khinchin's WLLN*)

Also

$$\begin{aligned}
\left|-\frac{1}{n} l_x'''(\theta')\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\ln f_\theta(X_i)\right|_{\theta'} &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3}{\partial\theta^3}\ln f_\theta(X_i)|_{\theta'}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} M(X_i) \text{ by condition } (vii) \\
&\longrightarrow E_{\theta_0} M(X_1) \text{ (by Khinchin's WLLN)} \\
&< \infty
\end{aligned}$$

And $\hat{\theta}_n - \theta_0 \xrightarrow{P} 0$ as $\hat{\theta}_n$ is consistent for θ.

$$\implies D_2 = -\frac{(\hat{\theta}_n - \theta_0)}{2}\frac{1}{n} l_x'''(\theta') \xrightarrow{P} 0 \quad \ldots \quad \ldots \quad (***)$$

$(**)$ and $(***) \implies D_1 + D_2 \xrightarrow{P} i(\theta_0)$

$\implies \frac{N}{D_1+D_2} \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)})$ under θ_0 [as by Slutsky's theorem
$X \xrightarrow{\mathcal{D}} Z$, $Y \xrightarrow{P} c \implies \frac{X}{Y} \xrightarrow{\mathcal{D}} \frac{Z}{c}$ provided Y and c are non-zero]
That is

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)}) \ \ldots \ \ldots \ (\odot)$$

Proved

Note: Any estimator satisfying ($\odot$) is called Efficient Likelihood Estimator (ELE).

Note: If the Likelihood equation has unique root, then the MLE is asymptotically efficient.

Again if the probability of multiroot for likelihood equation tends to 0, then also with probability tending to 1, the unique MLE is consistent and asymptotically efficient.

Note: The conditions under Theorem 3 are sufficient but not necessary for the consistency and asymptotic normality of a solution of the likelihood equation.

There are examples where not all the conditions are satisfied simultaneously; still a solution of the likelihood equation becomes consistant and asymptotically normal.

One such example is as follows:

Ex. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \theta)$, $\theta > 0$.

The likelihood equation of θ will be

$$L(\theta) = \frac{1}{(2\pi\theta)^{n/2}} \exp\left\{-\frac{1}{2\theta}\sum x_i^2\right\}$$

Then it is easy to check that the solution of the likelihood equation is $\hat{\theta}_n = \frac{1}{2\theta}\sum X_i^2$.

Also $\frac{X_i^2}{\theta} \sim \chi_1^2 \implies E\left(X_i^2\right) = \theta$, $V\left(X_i^2\right) = 2\theta^2$.

Hence from Lindeberg-Levy CLT we can write, as $n \longrightarrow \infty$,

$$\frac{\sum X_i^2 - n\theta}{\sqrt{2n\theta^2}} \xrightarrow{\mathcal{D}} N(0, 1).$$

e-Pathshala पाठशाला MHRD Govt. of India

i.e.

$$\frac{\sqrt{n}(\frac{\sum X_i^2}{n} - \theta)}{\sqrt{2\theta^2}} \xrightarrow{\mathcal{D}} N(0,1).$$

Or

$$\sqrt{n}(\frac{\sum X_i^2}{n} - \theta) \xrightarrow{\mathcal{D}} N(0, 2\theta^2).$$

Again $i(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i)\right] = E\left[-\frac{1}{2\theta^2} + \frac{X_i^2}{\theta^3}\right] = \frac{1}{2\theta^2}. \Longrightarrow$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta)}).$$

Hence asymptotic normality of the solution of likelihood equation holds.

Moreover we note that here, as $n \longrightarrow \infty$,

$$\hat{\theta}_n \xrightarrow{a.s.} \theta.$$

hence $\hat{\theta}_n$, the solution of the likelihood equation is consistant too.

But $\frac{\partial^3}{\partial\theta^3} \log f_\theta(x_i) = -\frac{1}{\theta^3} + \frac{3x^2}{\theta^4} \longrightarrow \infty$ as $\theta \longrightarrow 0$.

$\implies \frac{\partial^3}{\partial\theta^3} \log f_\theta(x_i)$ is not bounded in $0 < \theta < \infty$. And hence condition (vii) does not hold here.

The following theorem is a modification over Theorem 3 and it is so modified that the above cases are also covered.

Theorem 4. Suppose all the conditions of **Theorem 3.** remain unchanged except Condition (vii) which is now as stated below:

(vii) For every $\theta_0 \in \Omega$, $\exists \ \delta > 0$, a positive and twice differentiable function $g(\theta)$ and $M(x) \ni$

$$\left| \frac{\partial^2}{\partial \theta^2} [g(\theta) \ln f_\theta(x)] \right| \leq M(x) \ \forall \ \theta \in (\theta_0 - \delta, \theta_0 + \delta) \quad \forall \ x$$

with $E_{\theta_0} M(x) < \infty$.

Note that under $g(\theta) = 1$ above condition becomes identical with that of **Theorem 3.**

Proof. Referred to p. 500 of the book *Mathematical Methods of Statistics* by H. Cramer, and to G. Kulldorf (1957) *On the condition for consistency and asymptotic efficiency of maximum likelihood estimates*, *Skand. Aktusrietidskr.* **40**, 129 – 144.

In case of the above example if we choose $g(\theta) = \theta^3$ which is a positive and twice differentiable function of θ, $\left|\frac{\partial^2}{\partial\theta^2}\left[g(\theta)\ln f_\theta(x)\right]\right| = 1$. Taking $M(x) = 1 + x^2$, Condition (vii) above is satisfied with $E_{\theta_0}M(X) = 1 + \theta_0 < \infty$.

Note: The uniqueness of MLE does not ensure its asymptotic normality.

Ex. Let $X_1, \ldots, X_n$ be a random sample from $R(0, \theta)$ distribution. Then $X_{(n)}$ is the unique MLE of θ. But asymptotic distribution of $X_{(n)}$, by no way, is normal.

Ex.3 $f_\theta(x) = \exp\left[\theta T(x) + c(\theta) + D(x)\right].$

Let $\hat{\theta}_n$ be a solution of the likelihood equation
$E_\theta T(X) = \frac{1}{n}\sum T(X_i) = \overline{T}$, say.

$i(\theta) = E_\theta\left(\frac{\partial^2}{\partial\theta^2} lnf_\theta(X)\right) = -c''(\theta) = V_\theta T(X)$

So $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \frac{1}{V_{\theta_0} T(X)})$.

MLE based on Iteration : Some Typical Examples

Module 8

Saurav De

Department of Statistics
Presidency University

MLE based on Iteration : Some Typical Examples

Ex. (*Cauchy* $(\theta, 1)$ *Distribution*)

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ *Cauchy* $(\theta, 1)$; $\theta =$ population median (*location parameter*) and scale $= 1$. Then density is

$$f_\theta(x) = \frac{1}{\pi\left\{1 + (x - \theta)^2\right\}}$$

. Then the likelihood of θ is

$$L(\theta) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{1 + (x_i - \theta)^2}$$

$\implies$ the loglikelihood is : $l(\theta) =$ const $- \sum_{i=1}^{n} \log\left\{1 + (x_i - \theta)^2\right\}$

$\implies$ $l'(\theta) = 2\sum_{i=1}^{n} \frac{(x_i - \theta)}{1 + (x_i - \theta)^2}$.

MLE based on Iteration : Some Typical Examples

$\implies$ the likelihood equation is

$$\sum_{i=1}^{n} \frac{(x_i - \theta)}{1 + (x_i - \theta)^2} = 0 \quad \ldots \quad \ldots \quad (*)$$

In particular for $n = 2$, $(*) \implies$ a cubic equation;
for $n = 3$, $(*) \implies$ a 5-ic equation etc.

In general $(*) \implies$ a $(2n - 1)-$ic equation : practically impossible to solve for θ explicitly even for $n = 3$ or 4

Hence we call for an Iterative Method : A method to solve for θ or a mehtod to get MLE of θ

MLE based on Iteration : Some Typical Examples

Iterative Method : Denote MLE of θ by $\hat{\theta}_n$

To find $\hat{\theta}_n \ni \quad l'(\hat{\theta}_n) = 0$

Initial approximation $\hat{\theta}_{0n}$

Then from Taylor's expansion we can have

$$l'(\hat{\theta}_n) \approx l'(\hat{\theta}_{0n}) + (\hat{\theta}_n - \hat{\theta}_{0n}) l''(\hat{\theta}_{0n}).$$

provided $\hat{\theta}_{0n}$ is chosen in the close n.b.d. of $\hat{\theta}_n$.

$\Longrightarrow \quad \hat{\theta}_n \approx \hat{\theta}_{0n} - \frac{l'(\hat{\theta}_{0n})}{l''(\hat{\theta}_{0n})} \quad (\text{as } l'(\hat{\theta}_n) = 0)$

MLE based on Iteration : Some Typical Examples

$\Longrightarrow$

$$\hat{\theta}_{1n} = \hat{\theta}_{0n} - \frac{l'_{\mathbf{x}}(\hat{\theta}_{0n})}{l''_{\mathbf{x}}(\hat{\theta}_{0n})} \quad (\textit{1st approximation})$$

$$\hat{\theta}_{2n} = \hat{\theta}_{1n} - \frac{l'_{\mathbf{x}}(\hat{\theta}_{1n})}{l''_{\mathbf{x}}(\hat{\theta}_{1n})} \quad (\textit{2nd approximation})$$

and so on.

Stopping Rule: $\left\|\hat{\theta}_{r+1,n} - \hat{\theta}_{rn}\right\| < \epsilon$ for some r; ϵ : pre-assigned very small positive number

(*convergence criterion*)

Decision: $\hat{\theta}_{rn}$ or $\hat{\theta}_{r+1,n}$: Solution of the likelihood equation $(*)$ i.e. MLE of θ.

MLE based on Iteration : Some Typical Examples

Q. How to choose $\hat{\theta}_{0n}$ $\ni$ iterative estimator is better?

Def. T_n is called $\sqrt{n}-$consistent estimator for θ if $\sqrt{n}(T_n - \theta)$ is bounded in probability; i.e.
$\forall\ \epsilon > 0,\ \exists\ M$ and $n_0 \ni\ P\left\{\sqrt{n}\,|T_n - \theta| > M\right\} < \epsilon\ \forall\ n \geq n_0.$

In other words T_n is called $\sqrt{n}-$consistent estimator for θ if

$$\sqrt{n}(T_n - \theta) \xrightarrow{P} 0\,,$$

Note If $\sqrt{n}(T_n - \theta) \overset{\mathcal{D}}{\to} Z$ then T_n is $\sqrt{n}-$consistent for θ.

MLE based on Iteration : Some Typical Examples

Proof. For every $M > 0$

$$\lim P\left\{\sqrt{n}\,|T_n - \theta| > M\right\} = P\{|Z| > M\} = 1 - P\{|Z| \leq M\}$$

$= 1 - \{F_Z(M) - F_Z(-M)\} \downarrow \; M$, where $F_Z(\cdot)$ is the distribution function (DF) of Z.

Now for every $M > 0\,,\ P\{|Z| > M\} \leq \lim\limits_{M \longrightarrow \infty} P\{|Z| > M\} = 0$

$\Longrightarrow \lim \; P\left\{\sqrt{n}\,|T_n - \theta| > M\right\} = 0$ for every $M > 0$. Hence proved.

MLE based on Iteration : Some Typical Examples

Note : T_n is $\sqrt{n}$-consistent $\Longrightarrow$ T_n is consistent.

Proof. $|T_n - \theta| > M \Longrightarrow \sqrt{n}|T_n - \theta| > M$ for every $M > 0$.

Hence for every $M > 0$, $P[|T_n - \theta| > M] \leq P[\sqrt{n}|T_n - \theta| > M] \longrightarrow 0$ (*as* T_n *is* $\sqrt{n}$*-consistent for* θ)

$\Longrightarrow P[|T_n - \theta| > M] \longrightarrow 0$ for every $M > 0$.

i.e. T_n is consistent for θ. *Proved.*

MLE based on Iteration : Some Typical Examples

Result. If $\hat{\theta}_{0n}$ is $\sqrt{n}-$consistent for θ, the sequence of estimators

$$\hat{\theta}_{1n} = \hat{\theta}_{0n} - \frac{l_{\mathbf{x}}'(\hat{\theta}_{0n})}{l_{\mathbf{x}}''(\hat{\theta}_{0n})} \ ; \ \ n = 1, 2, \ldots$$

is asymptotically efficient under the assumptions of **Theorem 3.**

Proof : $\sqrt{n}(\hat{\theta}_{1n} - \theta_0) = \sqrt{n}(\hat{\theta}_{0n} - \theta_0) - \frac{\sqrt{n}\,\frac{1}{n} l_{\mathbf{x}}'(\hat{\theta}_{0n})}{\frac{1}{n} l_{\mathbf{x}}''(\hat{\theta}_{0n})} = U_n - V_n$ (say).

$$\text{Now } \frac{1}{n} l_{\mathbf{x}}'(\hat{\theta}_{0n}) = \frac{1}{n} \sum \frac{\partial}{\partial \theta} \log f_{\theta}(x_i)|_{\hat{\theta}_{0n}} \ ;$$

where $\frac{\partial}{\partial \theta} \log f_{\theta}(X_i)|_{\hat{\theta}_{0n}}$ is a sequence of iid random variables with $E_{\theta_0}\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X_i)\right) = 0$ and variance $= E_{\theta_0}\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X_i)\right)^2 = i(\theta_0)$.

MLE based on Iteration : Some Typical Examples

Hence by Lindeberg-Levy CLT

$$\sqrt{n}\,\frac{1}{n} l'_{\mathbf{x}}(\hat{\theta}_{0n}) \xrightarrow{\mathcal{D}} N(0, i(\theta_0))$$

$$\text{Again } \frac{1}{n} l''_{\mathbf{x}}(\hat{\theta}_{0n}) = \frac{1}{n}\sum \frac{\partial^2}{\partial\theta^2} \log f_\theta(x_i)|_{\hat{\theta}_{0n}}$$

where $\frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i)|_{\hat{\theta}_{0n}}$ is a sequence of iid random variables with $E_{\theta_0}\left(\frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i)\right) = -i(\theta_0)$. Hence by Khinchin's WLLN

$$\frac{1}{n} l''_{\mathbf{x}}(\hat{\theta}_{0n}) \xrightarrow{P} -i(\theta_0).$$

MLE based on Iteration : Some Typical Examples

So using Slutsky's Theorem, we have

$$V_n = \frac{\sqrt{n}\,\frac{1}{n} l'_{\mathbf{x}}(\hat{\theta}_{0n})}{\frac{1}{n} l''_{\mathbf{x}}(\hat{\theta}_{0n})} \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)})$$

Also as $\hat{\theta}_{0n}$ is $\sqrt{n}$-consistent for θ_0 , $U_n = \sqrt{n}(\hat{\theta}_{0n} - \theta_0) \xrightarrow{P} 0$. Hence again using Slutsky's Theorem finally we get

$$\sqrt{n}(\hat{\theta}_{1n} - \theta_0) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)}).$$

$\Longrightarrow$ $\hat{\theta}_{1n}$ is asymptotically efficient for θ_0. (*as* $\frac{1}{i(\theta_0)}$ *is the CRLB based on single random observation*)

Proved

MLE based on Iteration : Some Typical Examples

Rao's Scoring Method:

Result. Suppose $\hat{\theta}_{0n}$ is a $\sqrt{n}-$consistent estimator for θ. Then $\hat{\theta}_{1n} = \hat{\theta}_{0n} + \frac{l'_{\mathbf{x}}(\hat{\theta}_{0n})}{ni(\hat{\theta}_{0n})}$; is asymptotically efficient provided $i(\theta)$ is a continuous function of θ.

Proof. (Hint) $\frac{1}{n} l''_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} f_\theta(x_i)|_{\theta=\hat{\theta}_{0n}} \xrightarrow{P} -i(\theta_0)$ under θ_0 (*by Khinchin's WLLN*).

Again $i(\hat{\theta}_{0n}) \xrightarrow{P} i(\theta_0)$ under θ_0. Hence the proof is direct from the former **Result.**

MLE based on Iteration : Some Typical Examples

Now the density of Cauchy$(\theta, 1)$ is : $f_\theta(x) = \frac{1}{\pi\{1+(x-\theta)^2\}}$

$$\implies \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{2(x-\theta)}{1+(x-\theta)^2}$$

and

$$\begin{aligned}
i(\theta) = E\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right]^2 &= 4E\left[\frac{(x-\theta)^2}{(1+(x-\theta)^2)^2}\right] \\
&= \frac{8}{\pi}\int_0^\infty \frac{y^2}{(1+y^2)^3}dy \quad (y = x-\theta) \\
&= \frac{4}{\pi}\int_0^\infty \frac{z^{3/2-1}}{(1+z)^{3/2+3/2}}dz \quad (z = y^2) \\
&= \frac{4}{\pi}B(\frac{3}{2}, \frac{3}{2}) = \frac{1}{2}.
\end{aligned}$$

MLE based on Iteration : Some Typical Examples

Let $\tilde{X}_n$ = sample median.

Recall: If Z_p be the sample pth quantile

$$\frac{\sqrt{n}(Z_p - \zeta_p)}{\sqrt{\frac{p(1-p)}{f^2(\zeta_p)}}} \xrightarrow{\mathcal{D}} N(0,1)$$

as $n \longrightarrow \infty$ where f : population density and ζ_p : population pth quantile and n : sample size.

MLE based on Iteration : Some Typical Examples

Note that $\tilde{X}_n = Z_{\frac{1}{2}}$ and $\zeta_{\frac{1}{2}} = \theta$ for our Cauchy distribution.

$\implies p(1-p) = \frac{1}{4}$ and $f_\theta(\zeta_{\frac{1}{2}}) = \frac{1}{\pi}$ and hence

$$\frac{\sqrt{n}(\tilde{X}_n - \theta)}{\frac{\pi}{2}} \xrightarrow{\mathcal{D}} N(0,1)$$

Or

$$\sqrt{n}(\tilde{X}_n - \theta) \xrightarrow{\mathcal{D}} Z \sim N(0, \frac{\pi^2}{4})$$

$\implies \tilde{X}_n$ is a $\sqrt{n}-$consistent estimator of θ.

$\implies \tilde{X}_n$ can be taken as $\hat{\theta}_{0n}$.

MLE based on Iteration : Some Typical Examples

Now proceed as Rao's scoring method.

Illustration: Consider the following 15 sample values from a Cauchy $(\theta, 1)$ population

14.299	-0.165	-16.645	2.331	-0.108	122.423	1.952	0.372
-2.487	-0.925	-2.662	0.275	-5.111	-0.461	-2.561	

Here $\hat{\theta}_{0n} = \tilde{x} = -0.165$

$$
\begin{aligned}
\hat{\theta}_{1n} &= \hat{\theta}_{0n} + 2\sum_{i=1}^{15} \frac{(x_i - \hat{\theta}_{0n})}{\left\{1 + (x_i - \hat{\theta}_{0n})^2\right\}} \frac{1}{\frac{15}{2}} \\
&\approx -0.2779 \text{ correct up to 4 decimal places}
\end{aligned}
$$

MLE based on Iteration : Some Typical Examples

$$\implies \hat{\theta}_{2n} \approx -0.2916 \ , \ \hat{\theta}_{3n} \approx -0.2937, \ \hat{\theta}_{4n} \approx -0.2941, \ \hat{\theta}_{5n} \approx -0.2941 \ .$$

Here, $\hat{\theta}_{4n}$ and $\hat{\theta}_{5n}$ are both correct up to 4 decimal places (*from iterative recursion*).

$\implies$ the iteration converges correct upto 4 decimal places.

So the solution of the likelihood equation for θ, correct up to 4 decimal places is -0.2941.

MLE based on Iteration : Some Typical Examples

Computation using R :

(The solution of θ obtained in the Cauchy distribution example was implemented in this way in R.)

R Code and Output :

```
> samp=c
   (14.299,-0.165,-16.645,2.331,-0.108,122.423,1.952,0.372,-2.487,
+ -0.925,-2.662,0.275,-5.111,-0.461,-2.561)# sample drawn
> int_theta=median(samp)# consistent estimator of theta
> calculate=function(a,samp)
+ {
+ new_a=0
+ sum=0
+ for(i in 1:length(samp))
+ sum=sum+(((samp[i]-a)/(1+(samp[i]-a)^2))*(2/15))
+ new_a=a+(2*sum)
+ return(new_a)
+ }
```

MLE based on Iteration : Some Typical Examples

R Code and Output (continued) :

```
> i=0
> b=0
> new_theta=int_theta
> while(round(b,4)!=round(new_theta,4))
+ {
+ i=i+1
+ b=calculate(int_theta,samp)
+ cat("Value of theta in iteration",i," is : ",round(b,4),"\n")
+ if(b==int_theta)
+ break
+ else
+ {
+ new_theta=int_theta
+ int_theta=b
+ }
+ }
```

MLE based on Iteration : Some Typical Examples

R Code and Output (continued) :

```
Value of theta in iteration 1  is :  -0.2775
Value of theta in iteration 2  is :  -0.2916
Value of theta in iteration 3  is :  -0.2937
Value of theta in iteration 4  is :  -0.2941
Value of theta in iteration 5  is :  -0.2941
> cat("Converged value = ",round(b,4),"\n")
Converged value =  -0.2941
> cat("Number of iterations = ",i,"\n")
Number of iterations =  5
```

MLE based on Iteration : Some Typical Examples

Ex. (*Logistic Distribution*)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a Logistic distribution with common density

$$f_\theta(x) = \frac{\exp[-(x-\theta)]}{\{1+\exp[-(x-\theta)]\}^2} \; ; \quad -\infty < x < \infty \; , \quad -\infty < \theta < \infty.$$

$$\implies \; l_{\mathbf{x}}(\theta) = (-\sum x_i + n\theta) - 2\sum ln\left\{1 + \exp[-(x_i - \theta)]\right\}.$$

$$\implies \; l'_{\mathbf{x}}(\theta) = n - 2\sum \frac{1}{1+\exp[(x_i-\theta)]} \;\downarrow\; \theta$$

MLE based on Iteration : Some Typical Examples

Here $l''_{\mathbf{x}}(\theta) = -2\sum \frac{\exp[(x_i-\theta)]}{\{1+\exp[(x_i-\theta)]\}^2} < 0 \ \forall \ \theta.$

Also $l'_{\mathbf{x}}(\theta) \to -n$ as $\theta \to \infty$

And $l'_{\mathbf{x}}(\theta) \to n$ as $\theta \to -\infty$

$\implies \ \exists \ \hat{\theta}_{MLE} : l'_{\mathbf{x}}(\hat{\theta}_{MLE}) = 0$

i.e. $n - 2\sum \frac{1}{1+\exp[(x_i-\hat{\theta}_{MLE})]} = 0$

Now solve it by successive approx. method and get the unique $\hat{\theta}_{MLE}$ (*Like Cauchy*$(\theta, 1)$).

MLE based on Iteration : Some Typical Examples

Note : Under most of the truncated distributions, iterative method is the only option for finding ML estimates.

Application : Let $X_1, \ldots, X_n$ be a random sample drawn from a Poisson(λ) distribution where the mass point 0 is truncated. So the common pmf is

$$f_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{(1 - e^{-\lambda})\, x!} \;, \; x = 1, 2, \ldots \;; \; \lambda > 0.$$

The likelihood function of λ is

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{(1 - e^{-\lambda})^n \prod x!}$$

MLE based on Iteration : Some Typical Examples

$\implies$ the loglikelihood of λ is

$$l(\lambda) = \text{Const} \; - n\lambda + \sum x_i \ln \lambda - n \ln \left(1 - e^{-\lambda}\right)$$

So the likelihood equation $\frac{\partial}{\partial \lambda} l(\lambda) = 0 \implies$

$$\frac{\overline{x}}{\lambda} = \frac{e^{\lambda}}{e^{\lambda} - 1}$$

Obviously no explicit solution of the above equation is possible for λ.

Hence we have to depend on the Rao's scoring method or any other numerical method of iteration for finding ML estimate of λ.

MLE based on Iteration : Some Typical Examples

Multiparameter Case :

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)'$

Theorem 5. Under suitable assumptions with probability tending to 1 as $n \to \infty$; $\exists$ a root $\hat{\boldsymbol{\theta}}_n \ni$

(i) $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$

(ii) $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N_s(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$; $I_{jk}(\boldsymbol{\theta}) = E\left(-\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log f_{\boldsymbol{\theta}}(X)\right)$.

MLE based on Iteration : Some Typical Examples

Ex. Weibull density $f_{\alpha,\beta}(x) = \frac{\alpha}{\beta}x^{\alpha-1}\exp(-x^{\alpha}/\beta)\,;\ \alpha,\beta,x > 0.$

$l_{\mathbf{x}}(\theta) = n\log\alpha - n\log\beta + (\alpha-1)\sum\log x_i - \frac{1}{\beta}\sum x_i^{\alpha}$; based on a sample of size n.

$$\frac{\partial}{\partial\beta}l_{\mathbf{x}}(\theta) = 0 \Rightarrow -\frac{n}{\beta} + \frac{\sum x_i^{\alpha}}{\beta^2} = 0 \Leftrightarrow \hat{\beta} = \frac{\sum x_i^{\alpha}}{n}.$$

$$\frac{\partial}{\partial\alpha}l_{\mathbf{x}}(\theta) = 0 \Leftrightarrow \frac{n}{\alpha} + \sum\log x_i - \frac{\sum x_i^{\alpha}\log x_i}{\beta} = 0.$$

$$\Longrightarrow \frac{\sum x_i^{\alpha}\log x_i}{\sum x_i^{\alpha}} - \frac{1}{\alpha} = \frac{1}{n}\sum\log x_i$$

MLE based on Iteration : Some Typical Examples

Let $\phi(\alpha) = \frac{1}{n}\sum \log x_i - \left(\frac{\sum x_i^{\alpha} \log x_i}{\sum x_i^{\alpha}} - \frac{1}{\alpha} \right) = \frac{1}{n}\frac{\partial}{\partial \alpha} l_{\mathbf{x}}(\theta)$

Now check that

- $\phi'(\alpha) < 0$; i.e. $\phi(\alpha)$ is strictly decresing in α. (*you can use Cauchy-Swartz inequality to get it*)
- $\phi(\alpha) \longrightarrow \infty$ as $\alpha \to 0^+$ and $\phi(\alpha) \longrightarrow \frac{1}{n}\sum \log x_i - \log x_{(n)} < 0$ as $\alpha \to \infty$. (*you can use* $\frac{\sum x_i^{\alpha} \log x_i}{\sum x_i^{\alpha}} \leq \log x_{(n)}$ *and* $\frac{1}{n}\sum \log x_i \leq \log x_{(n)}$)

MLE based on Iteration : Some Typical Examples

Hence $\exists$ a unique $\alpha = \hat{\alpha} \ni \phi(\alpha) = 0 \Rightarrow \frac{\partial}{\partial \alpha} l_{\mathbf{x}}(\theta) = 0$. Also (*using Cauchy-Swartz inequality or something else*) check that the Hessian matrix

$$H = \begin{pmatrix} \frac{\partial^2}{\partial \alpha^2} l_{\mathbf{x}}(\theta)|_{\hat{\alpha},\hat{\beta}} & \frac{\partial^2}{\partial \alpha \partial \beta} l_{\mathbf{x}}(\theta)|_{\hat{\alpha},\hat{\beta}} \\ \frac{\partial^2}{\partial \alpha \partial \beta} l_{\mathbf{x}}(\theta)|_{\hat{\alpha},\hat{\beta}} & \frac{\partial^2}{\partial \beta^2} l_{\mathbf{x}}(\theta)|_{\hat{\alpha},\hat{\beta}} \end{pmatrix}$$

is negative definite (*as* $|H| > 0$ *and both the principal diagonal elements are negative*). As a result $\hat{\alpha}$ and $\hat{\beta}$ are the ML estimates of α and β.

MLE based on Iteration : Some Typical Examples

Application. The following data shows the maximum 24 Hr precifitation labels (in inches) recorded for 36 stalls in the period from 1990 to 1969 during the time they were over the mountains.

31.00	2.82	3.98	4.02	9.50	4.50	11.40	10.71	6.31	4.95	5.64
5.51	13.40	9.72	6.47	10.16	4.21	11.60	4.75	6.85	6.25	3.42
11.80	0.80	3.69	3.10	22.22	7.43	5.00	4.58	4.46	8.00	3.73
3.50	6.20	0.67								

a. Draw the histogram.

b. Identify the probability distribution.

c. Estimate the parameters using Method of Moments and the Method of ML.

d. Draw the fitted model.

MLE based on Iteration : Some Typical Examples

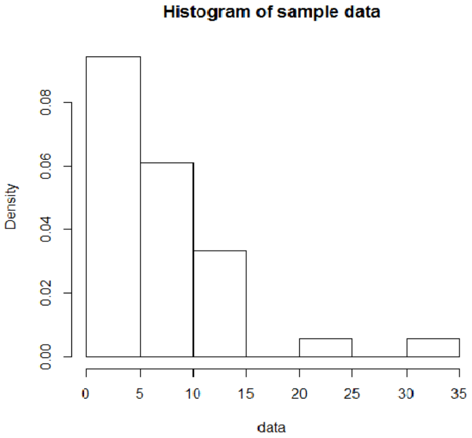


Figure: Histogram of the sample data provided.

MLE based on Iteration : Some Typical Examples

Discussion. From the histogram the underlying probability distribution can be considered to be Gamma with the common p.d.f.

$$f_{\alpha,\beta}(x) = \frac{1}{\beta^{\alpha}\Gamma\alpha} \exp\left(-\frac{x}{\beta}\right) x^{\alpha-1} \ , \ x, \alpha, \beta > 0$$

Then the likelihood function is

$$L(\alpha,\beta) = \frac{1}{\beta^{n\alpha}(\Gamma\alpha)^{n}} \exp\left(-\frac{\sum x_i}{\beta}\right) \prod x_i^{\alpha-1}$$

$$\Longrightarrow l(\alpha,\beta) = -n\alpha \ln\beta - n \ln\Gamma\alpha - \frac{\sum x_i}{\beta} + (\alpha-1)\sum \ln x_i$$

MLE based on Iteration : Some Typical Examples

Then

$$\frac{\partial}{\partial \beta} l(\alpha, \beta) = 0 \iff \hat{\beta} = \frac{\sum x_i}{n\hat{\alpha}} \quad \ldots \quad (*).$$

Now $\frac{\partial}{\partial \alpha} l(\alpha, \beta) = 0$ alongwith $(*)$ $\iff -n\frac{\Gamma\alpha'}{\Gamma\alpha} - n \ln \frac{\overline{x}}{\alpha} + \sum \ln x_i = 0$,

where $\Gamma\alpha' = \int_0^\infty \exp(-x) x^{\alpha-1} \log x \; dx$.

MLE based on Iteration : Some Typical Examples

- *Moment Method of Estimation*

$E(X_1) = \alpha\beta$ and $V(X_1) = \alpha\beta^2$. Hence using Moment Method the moment equations are

$\alpha\beta = \overline{X}$ and $\alpha\beta^2 = S^2$

where $\overline{X}$ and S^2 are the sample mean and sample variance respectively. On solving these equations the Moment Estimators of α and β are $\frac{\overline{X}^2}{S^2}$ and $\frac{S^2}{\overline{X}}$ respectively.

MLE based on Iteration : Some Typical Examples

- *ML Method of Estimation*

Using the Moment estimates as the initial approximation, one can solve the likelihood equations so obtained for α and β in the iterative way. The solution will give the ML estimates of α and β.

Finally replacing α and β by their estimates (*preferably ML estimates*) in the density model of the Gamma distribution, one can get the fitted model.

MLE based on Iteration : Some Typical Examples

Computation using R :

(The MLE and MME of α and β have been obtained in this way in R.)

R Code and Output :

```
> library(rGammaGamma)
> data=c(31,2.82,3.98,4.02,9.50,4.5,11.4,10.71,6.31,4.95,5.64,
+ 5.51,13.4,9.72,6.47,10.16,4.21,11.6,4.75,6.85,6.25,3.42,11.8,
+ 0.8,3.69,3.1,22.22,7.43,5,4.58,4.46,8,3.73,3.5,6.2,0.67)
> hist(data,freq=F,main="Histogram of sample data")
> X_bar=mean(data)
> S_2=var(data)
> alphaMME=(X_bar^2)/S_2
> betaMME=S_2/X_bar
> alphaMME# MME of shape
[1] 1.589584
> betaMME# MME of scale
[1] 4.584532
> gammaMLE(data)# MLE of shape and scale
   shape    scale
2.323328 3.143002
```

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Module 9

Saurav De

Department of Statistics
Presidency University

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Let $X_1, X_2, \ldots X_n$ be iid with common p.m.f. or p.d.f. $f_\theta(x)$.

$L_x(\theta)$ = Likelihood function of θ.

Consider the problem of testing $H : \theta \in \Omega_H$ against $K : \theta \in \Omega_K (\subseteq \Omega - \Omega_H)$; where Ω : parameter space and Ω_H : Parameter space under H.

The likelihood ratio(LR) criterion is defined as $\lambda(x) = \dfrac{\sup\limits_{\theta \in \Omega_H} L_x(\theta)}{\sup\limits_{\theta \in \Omega_H \cup \Omega_K} L_x(\theta)}$.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Note: $0 \leq \sup_{\theta \in \Omega_H} L_x(\theta) \leq \sup_{\theta \in \Omega_H \cup \Omega_K} L_x(\theta)$

i.e. $0 \leq \lambda(x) \leq 1$.

[An alternative form of LR criterion: $\lambda_1(x) = \dfrac{\sup_{\theta \in \Omega_H} L_x(\theta)}{\sup_{\theta \in \Omega_K} L_x(\theta)}$.

Naturally here $0 \leq \lambda_1(x) < \infty$. This form is seldom used; may be due to the fact that here the LR criterion is unbounded above]

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

A Discussion:

1. H is true $\implies$ numerator and denominator of $\lambda(x)$ will coincide.

   So $\lambda(x) = 1$ should imply that H is true trivially.

   Even a high value (*close to 1*) of $\lambda(x)$: evidence in favour of acceptance of H.

e-Pathshala पाठशाला MHRD Govt. of India
A Gateway to All Post Graduate Courses

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

On the other hand

2. H is not true $\implies$ the denominator of $\lambda(x)$ will give the supremum value of $L_x(\theta)$ because in that case • the most likely value of θ, if exists, will lie within Ω_K and hence • within $\Omega_H \cup \Omega_K$ but • not within Ω_H.

   $\implies$ the numerator will be significantly less compared to denominator.

   $\implies$ $\lambda(x)$ will be significantly low (close to 0).

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

e-Pathshala पाठशाला
MHRD Govt. of India

This discussion can easily motivate us to frame the critical region as follows.

Critical region: $\lambda(x) < c$, where c is $\ni$ size of the test is α.

[*or* $\lambda_1(x) < c_1$, *if the LR criterion is* $\lambda_1(x)$]

Note. If the distribution of $\lambda(x)$ is discrete, randomised test may be used.

Note. LRT entertains any kind of null and alternative hypotheses; simple as well as composite.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Note. Under simple null versus simple alternative hypothesis, LRT $\Longleftrightarrow$ Most Powerful Test using NP Lemma

Proof. Let $H : \theta = \theta_0$(known) versus $K : \theta = \theta_1$(known) [*i.e. Simple null versus simple alternative*]

Let $X_1, \ldots, X_n \sim$ pmf or pdf $f_\theta(x)$

$$\Longrightarrow \sup_{\theta=\theta_0} L(\theta) = f_{\theta_0}(\mathbf{x}) \quad \text{and} \quad \sup_{\theta=\theta_1} L(\theta) = f_{\theta_1}(\mathbf{x})$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

The LRT $\Longrightarrow$

$$\lambda_1(\mathbf{x}) = \frac{\sup\limits_{\theta=\theta_0} L(\theta)}{\sup\limits_{\theta=\theta_1} L(\theta)} < c$$

$$\Longleftrightarrow \frac{f_{\theta_0}(\mathbf{x})}{f_{\theta_1}(\mathbf{x})} < c$$

$$\Longleftrightarrow \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} > k \left(= \frac{1}{c}\right)$$

$\longrightarrow$ the MP Test from NP Lemma.

Proved

Note. If $\exists$ a sufficient statistic $T(X)$ based on X, then $\lambda(x)$ is a function of $T(X)$.

Proof. Using Neyman-Fisher Factorisation Theorem, we can write

$$\lambda(x) = \frac{\sup_{\theta \in \Omega_H} L_x(\theta)}{\sup_{\theta \in \Omega_H \cup \Omega_K} L_x(\theta)} = \frac{h(\mathbf{x}) \sup_{\theta \in \Omega_H} g_\theta(T(\mathbf{x}))}{h(\mathbf{x}) \sup_{\theta \in \Omega_H \cup \Omega_K} g_\theta(T(\mathbf{x}))} = u(T(\mathbf{x}));$$

a function of T($\mathbf{x}$). Hence proved.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Ex. Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ independently.

To test $H : \mu = 0$ versus (a) $K_1 : \mu \neq 0$ and (b) $K_2 : \mu > 0$.

$$L_x(\theta) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \mu)^2/2\sigma^2\right]$$

(a) $\Omega_H = \{(\mu, \sigma) : \mu = 0, \sigma > 0\}$; $\Omega_H \cup \Omega_{K_1} = \{(\mu, \sigma) : \mu \in \mathcal{R}, \sigma > 0\}$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Under $\Omega_H \cup \Omega_{K_1}$; $\hat{\mu} = \overline{x}$; $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Under Ω_H; $\hat{\sigma^2_H} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

LR criterion

$$\lambda(x) = \left(\frac{\hat{\sigma^2}}{\hat{\sigma^2_H}}\right)^{n/2} = \left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2}\right)^{n/2} = \left(1 + \frac{n\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^{-n/2}.$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Critical region:

$$\lambda(x) < c \iff \left(1 + \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{-n/2} < c \iff \frac{\sqrt{n}|\bar{x}|}{\sqrt{\sum(x_i - \bar{x})^2}} > c$$

$$\iff \frac{\sqrt{n}|\bar{x}|}{s} > t_{\alpha/2, n-1} \quad \text{where} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This is UMPU test with test statistic $t_H = \frac{\sqrt{n}|\bar{x}|}{s}$ which, under H, follows t-distribution with $n-1$ degrees of freedom.

(b) $\Omega_H \cup \Omega_{K_2} = \{(\mu, \sigma) : \mu \geq 0, \sigma > 0\}$

$$\begin{aligned} \hat{\mu} &= \overline{x} \text{ if } \overline{x} \geq 0 \\ &= 0 \text{ if } \overline{x} < 0 \end{aligned}$$

So $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$

$$\begin{aligned} \lambda(x) &= 1 \qquad , \overline{x} < 0 \text{ (case of trivial acceptence of } H) \\ &= \left(1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n\overline{x}^2}\right)^{n/2} \qquad , \overline{x} \geq 0 \end{aligned}$$

$\lambda(x) < c \iff \frac{\sqrt{n}\ \overline{x}}{s} > t_{\alpha, n-1}$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Ex. Let $X_1, X_2, \ldots, X_n$ be n independent bernoulli(p) variables. Suppose we are to test

$$H : p \leq p_0 \text{ versus } K : p > p_0.$$

Here $\Omega_H = \{p : 0 \leq p \leq p_0\}$ and $\Omega_H \cup \Omega_K = \{p : 0 \leq p \leq 1\}$.

Under $\Omega_H \cup \Omega_K$, the ML estimate of p is the sample mean $\overline{x}$ and hence the supremum of the likelihood is

$$\sup_{\Omega_H \cup \Omega_K} L(p) = (\overline{x})^{n\overline{x}}(1-\overline{x})^{n(1-\overline{x})}.$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Under Ω_H i.e. under $p \leq p_0$, MLE of p is

$$\begin{aligned} \hat{p} &= \overline{x} \ , \ \overline{x} \leq p_0 \\ &= p_0 \ , \ \overline{x} > p_0 \ \ (\textit{Restricted MLE of } p) \end{aligned}$$

$$\begin{aligned} \text{Hence } \sup_{\Omega_H} L(p) &= (\overline{x})^{n\overline{x}}(1-\overline{x})^{n(1-\overline{x})} \ , \ \overline{x} \leq p_0 \\ &= (p_0)^{n\overline{x}}(1-p_0)^{n(1-\overline{x})} \ , \ \overline{x} > p_0 \end{aligned}$$

$$\begin{aligned} \Longrightarrow \lambda(\mathbf{x}) &= 1 \ , \ \overline{x} \leq p_0 \\ &= \frac{(p_0)^{n\overline{x}}(1-p_0)^{n(1-\overline{x})}}{(\overline{x})^{n\overline{x}}(1-\overline{x})^{n(1-\overline{x})}} \ , \ \overline{x} \leq p_0 \end{aligned}$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

So $\lambda(\mathbf{x}) = 1$ for $\overline{x} \le p_0$ but $\lambda(\mathbf{x}) = \frac{(p_0)^{n\overline{x}}(1-p_0)^{n(1-\overline{x})}}{(\overline{x})^{n\overline{x}}(1-\overline{x})^{n(1-\overline{x})}} \le 1$ for $\overline{x} > p_0$.

$\implies \lambda(\mathbf{x}) \downarrow \overline{x}$

Thus the LRT critical region $\lambda(\mathbf{x}) < c \iff \overline{x} > c_1$ or equivalently $\sum x_i > c_2$, where c_2 is $\ni$

$$\sup_{H} P_p\left[\sum X_i > c_2\right] \le \alpha \text{ i.e. } \sup_{p \le p_0} P_p\left[\sum X_i > c_2\right] \le \alpha$$

, α being the given level of significance.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

In this case $\sum X_i \sim Bin(n, p)$ distribution.

Now we know $P_p[\sum X_i > k] = I_p(n-k, k+1)$, where

$$I_p(n-k, k+1) = (B(n-k, k+1))^{-1} \int_0^p u^{n-k-1}(1-u)^k du$$

$B(n-k, k+1)$ being the Beta integral and $I_p(n-k, k+1)$, the incomplete Beta function.

As $I_p(n-k, k+1) \uparrow p$ (*evident from the definition of* $I_p(n-k, k+1)$)

$$\implies \sup_{p \le p_0} P_p\left[\sum X_i > c_2\right] = P_{p_0}\left[\sum X_i > c_2\right] \le \alpha.$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Thus the LRT for testing $H : p \leq p_0$ against $K : p > p_0$ is

Reject H if $\sum X_i > c_2$, where c_2 is the smallest integer $\ni$

$$P_{p_0}\left[\sum X_i > c_2\right] \leq \alpha.$$

Note : In this case the LR test coincides with the UMP test.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Ex. Let $X_1, X_2, \ldots, X_n \sim R(0, \theta)$ independently. Then

$$\begin{aligned} L_x(\theta) &= \frac{1}{\theta^n} \text{ if } x_{(n)} \leq \theta \\ &= 0 \text{ o.w.} \end{aligned}$$

Let $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$.

Under $\Omega_H \cup \Omega_K$; $\hat{\theta} = X_{(n)}$. Hence the LR criterion is

$$\lambda(x) = \frac{\sup_{\theta \in \Omega_H} L_x(\theta)}{\sup_{\theta \in \Omega_H \cup \Omega_K} L_x(\theta)} = \frac{\frac{1}{\theta_0^n} I_x(x_{(n)}, \theta_0)}{\frac{1}{x_{(n)}^n}} \begin{aligned}[t] &= 0 \text{ if } x_{(n)} > \theta_0 \\ &\quad \text{(case of trivial rejection of } H \\ &= \left(\frac{x_{(n)}}{\theta_0}\right)^n \text{ if } x_{(n)} \leq \theta_0 \end{aligned}$$

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Now the LR test: $0 \leq \lambda(x) < c$

where c is determined from size-α condition.

$\iff$ LR test: $x_{(n)} < d$ or $x_{(n)} > \theta_0$
where d is determined from size-α condition.

$\implies$ LR test: $x_{(n)} < \theta_0 \ \alpha^{1/n}$ or $x_{(n)} > \theta_0$.

This is an UMP size-α test.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Ex. A random sample of size n is taken from the p.m.f.
$P(X_j = x_j) = p_j,\ j = 1, 2, 3, 4,\ 0 < p_j < 1$. Find the form of LR test of
$H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ against
$H_1 : p_1 = p_2 = p/2\ ,\ p_3 = p_4 = (1 - p)/2,\ 0 < p < 1.$

Let $n_j = \#$ times the value x_j appears in the sample of size n (fixed).

Obviously $\sum_{i=1}^{4} n_j = n.$

Also $\mathbf{n} = (n_1, n_2, n_3, n_4)' \sim MN(n; p_1, p_2, p_3, p_4)\ ,\ \sum_{i=1}^{4} p_j = 1.$

Then under H_0 the likelihood function is

$L_{H_0}(\mathbf{p}|\mathbf{n}) = C(\mathbf{n}) \cdot \left(\frac{1}{4}\right)^n$; a constant, where $C(\mathbf{n}) = \frac{n!}{n_1!n_2!n_3!n_4!}$

Similarly $L_{H_1}(\mathbf{p}|\mathbf{n}) = C(\mathbf{n}) \cdot p^t(1-p)^{n-t}$ where $t = n_1 + n_2$.

Now it is not difficult to get that the maximum of $L_{H_1}(\mathbf{p}|\mathbf{n})$ is $\frac{C(\mathbf{n})}{n^n} \cdot t^t(n-t)^{n-t}$ which attains at $p = \frac{t}{n}$.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Now the LR : $\lambda(\mathbf{n}) = \dfrac{C(\mathbf{n}) \cdot \left(\frac{1}{4}\right)^n}{\frac{C(\mathbf{n})}{n^n} \cdot t^t (n-t)^{n-t}}$

$\implies$ the critical region based on the LR criterion is $\{\lambda(\mathbf{n})) < K_1\}$

$\iff \left\{ \dfrac{\left(\frac{n}{4}\right)^n}{t^t (n-t)^{n-t}} < K_1 \right\}$

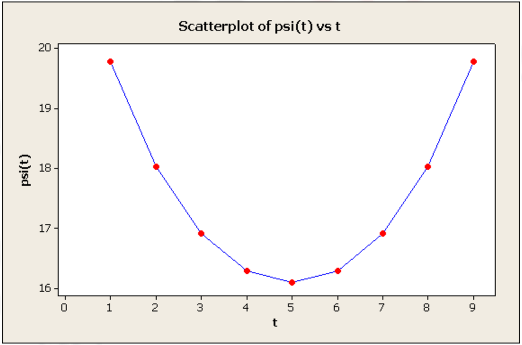
$\iff \{\psi(t) > K_2\}$

where $\psi(t) = t \ln t + (n-t) \ln(n-t)$, and K_1 and K_2 are suitable constants.

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

Now check that $\psi(t)$ is minimum at $t = \frac{n}{2}$ and $\psi(\frac{n}{2} - t) = \psi(\frac{n}{2} + t)\ \forall t$.
This is evident also from the following graph:

LIKELIHOOD RATIO TEST (LRT): Basic Ideas



LIKELIHOOD RATIO TEST (LRT): Basic Ideas

So $\{\psi(t) > K_2\} \iff \{t < K \text{ or } t > n - K\}$

where the constant K is $\ni$

$P_{H_0}[T < K] \leq \alpha/2$ but $P_{H_0}[T < K + 1] > \alpha/2$

with $T = n_1 + n_2 \overset{H_0}{\sim}$ Bin $(n, 1/2)$ (*from the marginal distribution of multinomial distribution*).

LIKELIHOOD RATIO TEST (LRT): Basic Ideas

TRY YOURSELF!

M9.1. Based on a random sample of size n from a Poisson (λ) distribution, give an LRT for testing (a) $H : \lambda = 2$ versus $K_1 : \lambda \neq 2$ and (b) $H : \lambda \geq 2$ versus $K_2 : \lambda < 2$

M9.2. A die is tossed 60 times in order to test
$H : P\{j\} = 1/6\,,\, j = 1, 2, \ldots, 6$ (i.e. die is fair) against
$K : P\{2j-1\} = 1/9\,,\, P\{2j\} = 2/9\,,\, j = 1, 2, 3.$
Provide the LR test.

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 9 . . .

M9. 1. In part (a), $\Omega_H = \{\lambda = 2\}$, a singleton set and $\Omega_H \cup \Omega_{K_1} = \{\lambda : 0 \le \lambda < \infty\}$; the unrestricted parameter space of λ of which the sample mean $\overline{x}$ is the MLE.

In part (b), under $\Omega_H = \{\lambda \ge 2\}$; a restricted parameter space of λ, the MLE of λ is

$$\begin{aligned} \hat{\lambda} &= \overline{x} \, , \; \overline{x} \ge 2 \\ &= 2 \, , \; \overline{x} < 2 \end{aligned}$$

Now proceed straightway as discussed in the worked-out examples.

M9. 2. The solution is very similar but comparatively little easier to the solution of the last worked-out example in this module.

e-Pathshala पाठशाला MHRD Govt. of India

Some Further Properties of LRT

Module 10

Saurav De

Department of Statistics
Presidency University

LRT may be worthless :

Suppose the pmf of X with parameter θ is as follows:

$$
\begin{aligned}
p_{\theta=0}(x) &= \alpha \ , \ x = 0 \\
&= \frac{1}{2} - \alpha \, , \ x = \pm 1 \\
&= \frac{\alpha}{2} \ , \ x = \pm 2 \\
&= 0 \text{ otherwise}
\end{aligned}
$$

$$
\begin{aligned}
p_{\theta\in(0,1]}(x) &= \frac{1-c}{1-\alpha}\alpha \ \text{ if } \ x=0 \\
&= \frac{1-c}{1-\alpha}(\frac{1}{2}-\alpha) \ \text{ if } \ x=\pm 1 \\
&= c\theta \ \text{ if } \ x=2 \ , \ 0<\alpha<1/2 \\
&= c(1-\theta) \ \text{ if } \ x=-2 \ , \ \alpha/(2-\alpha)<c<\alpha \\
&= 0 \ \text{ otherwise}
\end{aligned}
$$

To test $H:\theta=0$ against $K:\theta>0$.

The LR criterion $\lambda(x)=\frac{L_x(\theta=0)}{\sup\limits_{0\le\theta\le 1} L_x(\theta)}$

Some Further Properties of LRT

$$\implies \lambda(0) = \frac{\alpha}{\max\left\{\alpha, \left(\frac{1-c}{1-\alpha}\right)\alpha\right\}}$$

Given $c < \alpha \implies 1 - c > 1 - \alpha$ i.e. $\frac{1-c}{1-\alpha} > 1 \iff \left(\frac{1-c}{1-\alpha}\right)\alpha > \alpha.$

As a result $\lambda(0) = \frac{\alpha}{\left(\frac{1-c}{1-\alpha}\right)\alpha} = \frac{1-\alpha}{1-c}$

Similarly $\lambda(1) = \frac{1/2-\alpha}{\max\left\{(1/2-\alpha), \left(\frac{1-c}{1-\alpha}\right)(1/2-\alpha)\right\}} = \frac{1-\alpha}{1-c} = \lambda(-1)$

In the same way

$$\lambda(2) = \frac{\frac{\alpha}{2}}{\sup\left\{\frac{\alpha}{2}, \sup\limits_{0<\theta\leq 1} c\,\theta\right\}} = \frac{\frac{\alpha}{2}}{\sup\left\{\frac{\alpha}{2}, c\right\}}$$

Now $\frac{\alpha}{2-\alpha} < c \iff \frac{2-\alpha}{\alpha} > \frac{1}{c}$ or $\frac{2}{\alpha} - 1 > \frac{1}{c}$

$$\Longrightarrow \frac{2}{\alpha} > \frac{1}{c} \iff \frac{\alpha}{2} < c$$

$\Longrightarrow \lambda(2) = \frac{\frac{\alpha}{2}}{c} = \frac{\alpha}{2c}$. Similarly we get $\lambda(-2) = \frac{\alpha}{2c}$.

Some Further Properties of LRT

Again $\frac{\alpha}{2-\alpha} < c \implies \frac{2}{\alpha} - 1 > \frac{1}{c}$

i.e. $\frac{2}{\alpha} - 2 > \frac{1}{c} - 1 = \frac{1-c}{c}$

or $2\left(\frac{1-\alpha}{\alpha}\right) > \frac{1-c}{c} \iff \frac{1-\alpha}{1-c} > \frac{\alpha}{2c}$

hence the LR test: $\lambda(x) < d \iff \lambda(x) \leq \frac{\alpha}{2c}$

Reason : $\lambda(x)$ has only two positive finite values $\frac{1-\alpha}{1-c}$ and $\frac{\alpha}{2c}$ and among these two, the smaller is $\frac{\alpha}{2c}$.

Some Further Properties of LRT

But $\lambda(x) \leq \frac{\alpha}{2c} \iff \lambda(x) = \frac{\alpha}{2c} \iff x = \pm 2.$

$\implies$ the dritical region of LRT : $\{x = \pm 2\}$.

Size of the test $= P_{\theta=0}[x = \pm 2] = 2 \times \frac{\alpha}{2} = \alpha.$

Power function is $P_{\theta}[x = \pm 2] = c\theta + c(1 - \theta) = c < \alpha \ \forall \ \theta > 0.$

$\implies$ the LR test is biased.

e-Pathshala पाठशाला MHRD Govt. of India

Some Further Properties of LRT

Define a test function $\phi^*(x) \equiv \alpha$ (i.e. constant) $\forall\ x$

$\longrightarrow$ a trivial size α test.

The power function is $E_\theta\, \phi^*(x) = \alpha\ \forall\ \theta \geq 0$.

$\Longrightarrow$ ϕ^* is **unbiased** and is **uniformly more powerful** than the above LRT, though ϕ^* is a trivial test.

$\Longrightarrow$ the above LRT is practically worthless.

Large sample properties of an LRT :

Consider the following theorems.

Theorem under single parameter case : Consider the problem of testing $H : \theta = \theta_0$ against $K : \theta \neq \theta_0$. Under the assumption of Theorem 3,

$$-2 \log \lambda(\mathbf{x}) \xrightarrow{\mathcal{D}} \chi^2_1 \,, \text{ under H}$$

Proof. By definition

$\lambda(\mathbf{x}) = \frac{L(\theta_0)}{L(\hat{\theta}_n)}$, $\hat{\theta}_n$ denoting MLE of θ.

Or $\log \lambda(\mathbf{x}) = l(\theta_0) - l(\hat{\theta}_n)$ ($l(\theta)$: the loglikelihood function of θ)

Expanding $l(\theta_0)$ w.r.t. $\hat{\theta}_n$ by Taylor's expansion $\Longrightarrow$

$$l(\theta_0) \approx l(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)l'(\hat{\theta}_n) + \frac{(\hat{\theta}_n - \theta_0)^2}{2} l''(\hat{\theta}_n)$$

$$l(\hat{\theta}_n) - l(\theta_0) \approx \frac{(\hat{\theta}_n - \theta_0)^2}{2}\left(-l''(\hat{\theta}_n)\right)$$

(as $l'(\hat{\theta}_n)$ vanishes because $\hat{\theta}_n$ is MLE)

Moreover $-\frac{1}{n}l''(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2} f_\theta(x_i) \mid_{\theta=\hat{\theta}_n}$ (also known as observed Fisher's Information) is asymptotically equal to $i(\hat{\theta}_n)$.

$$\implies 2\left[l(\hat{\theta}_n) - l(\theta_0)\right] \overset{a}{\approx} \left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right)^2 i(\hat{\theta}_n).$$

e-Pathshala पाठशाला MHRD Govt. of India

Thus $-2\log\lambda(\mathbf{x}) = 2\left[l(\hat{\theta}_n) - l(\theta_0)\right] \stackrel{a}{\approx} \left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right)^2 i(\hat{\theta}_n)$.

Asymptotic distribution of MLE $\Longrightarrow$

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)})$$

Also $i(\hat{\theta}_n) \longrightarrow i(\theta_0)$ under H. So

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\sqrt{i(\hat{\theta}_n)} \xrightarrow{\mathcal{D}} N(0,1) \text{ under H}$$

Hence the theorem.

Theorem on multiparameter case :

Suppose we are to test $H : \boldsymbol{\theta}_{s\times 1} = \boldsymbol{\theta}_0$ against $H : \boldsymbol{\theta}_{s\times 1} \neq \boldsymbol{\theta}_0$.

Under the assumption of Theorem 4,

$$-2\log\lambda(\mathbf{x}) \xrightarrow{\mathcal{D}} \chi^2_s\,, \text{ under H}$$

Proceed just like the earlier proof using asymptotic multivariate normality of the vector valued MLE $\hat{\boldsymbol{\theta}}_n$ of the $s-$component parameter vector $\boldsymbol{\theta}$.

Note on asymptotic distribution :

To test $H : \theta_i = g_i(\beta_1, \ldots, \beta_r)$; $i = 1, \ldots, s, r < s$; $\beta_1, \ldots, \beta_r$ being unknown,

$$-2 \log \lambda(\mathbf{x}) \xrightarrow{\mathcal{D}} \chi^2_{s-r} \text{ , under H}$$

where $s - r =$ No. of parameters specified by H.

Above result gets several applications on testing of hypotheses.

Application 1 : Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently.
To test $H : \mu = 0$ versus $k : \mu \neq 0$.

Here parameters are (μ, σ^2) i.e. $s = 2$.
σ remains unknown althrough , i.e. $r = 1$.
So here

$$-2 \log \lambda(\mathbf{x}) \xrightarrow{\mathcal{D}} \chi^2_1, \text{ under H.}$$

Some Further Properties of LRT

Application 2 : Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N_p(\boldsymbol{\mu}, \Sigma)$ independently.

To test $H : \Sigma = \sigma^2 I_p$ against $K : \Sigma \neq \sigma^2 I_p$ (known as Sphericity Test)

Here # parameters in general $= s = p + \frac{p(p+1)}{2}$ and # unknown parameters $= r = p + 1$.

$\implies$ # parameters specified by $H = s - r = \frac{p(p+1)}{2} - 1$.

$$\implies -2 \log \lambda(\mathbf{x}) \xrightarrow{\mathcal{D}} \chi^2_{\frac{p(p+1)}{2}-1}, \text{ under H.}$$

Consistency of LRT :

Consider the problem of testing $H : \theta \in \Omega_H$ against $K : \theta \in \Omega_K (\subseteq \Omega - \Omega_H)$.

Naturally LR criterion will be $\lambda(\mathbf{x}) = \frac{L_{\mathbf{x}}(\hat{\theta}_H)}{L_{\mathbf{x}}(\hat{\theta})}$

where $\hat{\theta}$: MLE of θ under Ω and $\hat{\theta}_H$: MLE of θ under Ω_H

Suppose $\hat{\theta}_H \xrightarrow{P} \theta \in \Omega_H$ under Ω_H and

$\hat{\theta} \xrightarrow{P} \theta \in \Omega_H \cup \Omega_K$ under Ω.

Some Further Properties of LRT

Let the LR test be : $\lambda(\mathbf{x}) < C_n(\alpha)$

where $C_n(\alpha) \ni$ the size of the test is α i.e.

$$\sup_{\theta \in \Omega_H} P_\theta \left[\lambda(\mathbf{X}) < C_n(\alpha)\right] = \alpha$$

Under $\theta \in \Omega_H$, $\lambda(\mathbf{x}) \xrightarrow{P} 1 \implies C_n(\alpha) \longrightarrow 1$ as $n \to \infty$.

ePathshala पाठशाला MHRD Govt. of India
A Gateway to All Post Graduate Courses

Some Further Properties of LRT

Also power at $\theta \in \Omega_K = P_\theta\left[\lambda(\mathbf{X}) < C_n(\alpha)\right]$

Under $\theta \in \Omega_K$; $\lambda(\mathbf{x}) \xrightarrow{P} \gamma < 1$
and $C_n(\alpha) \longrightarrow 1$ as $n \to \infty$.

$$\implies P_\theta\left[\lambda(\mathbf{X}) < C_n(\alpha)\right] \longrightarrow 1 \text{ as } n \to \infty \ \forall\ \theta \in \Omega_K$$

Since power $\longrightarrow 1$ as $n \to \infty$, the LR test is consistent.

LRT : Applications under Single Distributions

Module 11

Saurav De

Department of Statistics
Presidency University

MHRD
Govt. of India

LRT : Applications under Single Distributions

A common notion : *The randomised test is applicable only when the parent population is discrete.*

It is an incomplete notion !

A randomised test is called for whenever the LR criterion $\lambda(\mathbf{x})$ has discrete probability distribution, no matter the parent population is discrete or continuous.

LRT : Applications under Single Distributions

An example: Let $X_1, X_2 \sim R(\theta, \theta+1)$ independently. To test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, $\theta_0 < \theta_1 < \theta_0 + 1$

$\implies \mathcal{X}_0 = \{\theta_0 \leq x_i \leq \theta_0 + 1, i = 1, 2\}$: Sample space under H_0
and $\mathcal{X}_0 = \{\theta_1 \leq x_i \leq \theta_1 + 1, i = 1, 2\}$: Sample space under H_1.
So

$$\begin{aligned} L_{H_0}(\theta) &= 1\,,\ (x_1, x_2) \in \mathcal{X}_0 \\ &= 0\,,\ (x_1, x_2) \notin \mathcal{X}_0 \end{aligned}$$

and

$$\begin{aligned} L_{H_1}(\theta) &= 1\,,\ (x_1, x_2) \in \mathcal{X}_1 \\ &= 0\,,\ (x_1, x_2) \notin \mathcal{X}_1 \end{aligned}$$

LRT : Applications under Single Distributions

Now

$$\lambda(\mathbf{x}) = \frac{L_{H_0}(\theta)}{L_{H_1}(\theta)} \begin{array}{rl} = & \infty\,,\ \mathbf{x} \in \mathcal{X}_0 \cap \mathcal{X}_1^c \\ = & 1\,,\ \mathbf{x} \in \mathcal{X}_0 \cap \mathcal{X}_1 \\ = & 0\,,\ \mathbf{x} \in \mathcal{X}_0^c \cap \mathcal{X}_1 \end{array}$$

$\implies P_{H_0}[\lambda(\mathbf{x}) = \infty] = P_{H_0}[(X_1, X_2) \in \mathcal{X}_0 \cap \mathcal{X}_1^c] = 1 - (1 - \theta_1 + \theta_0)^2,$

$P_{H_0}[\lambda(\mathbf{x}) = 1] = P_{H_0}[(X_1, X_2) \in \mathcal{X}_0 \cap \mathcal{X}_1] = (1 - \theta_1 + \theta_0)^2$

and $P_{H_0}[\lambda(\mathbf{x}) = 0] = P_{H_0}[(X_1, X_2) \in \mathcal{X}_0^c \cap \mathcal{X}_1] = 0$

$\Longrightarrow \lambda(\mathbf{x})$ has a discrete probability distribution.

Also the critical region of LRT is

$W = \{\lambda(\mathbf{x}) < c\}$, c to be obtained from size condition.

Let given level of significance : α. Also

$$\begin{aligned} P_{H_0}[\lambda(\mathbf{x}) < c] &= P_{H_0}[\lambda(\mathbf{x}) = 0] = 0 \text{ if } 0 < c \leq 1 \\ &= P_{H_0}[\lambda(\mathbf{x}) \leq 1] = (1 - \theta_1 + \theta_0)^2 \text{ if } 1 < c < \infty \end{aligned}$$

LRT : Applications under Single Distributions

If $(1-\theta_1+\theta_0)^2 = \alpha$ i.e. if $\theta_1 - \theta_0 = 1-\alpha^{\frac{1}{2}}$, then the CR of level (as well as size) α is

$$W = \{\lambda(\mathbf{x}) < c\}\ ,\ \text{for any } c \in (1,\infty).$$

If $\alpha < (1-\theta_1+\theta_0)^2$, call for a randomised test

$$\begin{aligned}\phi(\mathbf{x}) &= 1\,,\ \mathbf{x} \in \mathcal{X}_0 \cap \mathcal{X}_1^c \\ &= \gamma_\alpha\,,\ \mathbf{x} \in \mathcal{X}_0 \cap \mathcal{X}_1 \\ &= 0\,,\ \mathbf{x} \in \mathcal{X}_0^c \cap \mathcal{X}_1\end{aligned}$$

where $\gamma_\alpha \ni\ E_{H_0}\phi(\mathbf{X}) = \alpha \iff \gamma_\alpha = \frac{\alpha}{(1-\theta_1+\theta_0)^2}$.

LRT : Applications under Single Distributions

Aliter: define a CR

$$W = \{c < x_i < \theta_0 + 1\,,\; i = 1, 2\,,\; \theta_0 < c\}$$

where c is $\ni$ $P_{H_0}(W) = \alpha$

$$\text{Now } P_{H_0}(W) = \left(\int_c^{\theta_0+1} dx \right)^2 = (1 - (c - \theta_0))^2$$

$\implies (1 - (c - \theta_0))^2 = \alpha$, from size condition.

$\implies$ on simplification $c = 1 + \theta_0 - \sqrt{\alpha}$.

LRT : Applications under Single Distributions

Application 1: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. To test $H : \sigma = \sigma_0$(known) versus $(a) K_1 : \sigma \neq \sigma_0$ and $(b) K_2 : \sigma > \sigma_0$.
Here $\theta = (\mu, \sigma)$.

$$\Omega_H = \{(\mu, \sigma) : \mu \in \mathcal{R}\,,\ \sigma = \sigma_0\}\,,$$

$$\Omega_{K_1} = \{(\mu, \sigma) : \mu \in \mathcal{R}\,,\ \sigma > 0\,,\ \sigma \neq \sigma_0\}$$

and $\Omega_{K_2} = \{(\mu, \sigma) : \mu \in \mathcal{R}\,,\ \sigma_0 < \sigma < \infty\}$. The likelihood is

$$L(\theta) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right\}$$

$$\Longrightarrow \max_{\theta \in \Omega_H} L(\theta) = (2\pi)^{-n/2} (\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2} \sum (x_i - \overline{x})^2\right\}$$

LRT : Applications under Single Distributions

i.e. $\max\limits_{\theta \in \Omega_H} L(\theta) = (2\pi)^{-n/2}(\sigma_0^2)^{-n/2} \exp\left\{-\frac{ns^2}{2\sigma_0^2}\right\}$, $s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$: sample variance

Again under $\Omega_H \cup \Omega_{K_1} = \{(\mu, \sigma) : \mu \in \mathcal{R}\,, \, \sigma > 0\}$, MLE of $\sigma^2 = s^2$.

$$\Longrightarrow \max_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) = (2\pi)^{-n/2}(s^2)^{-n/2} \exp\left\{-\frac{ns^2}{2s^2}\right\} = \frac{\exp\{-n/2\}}{(2\pi s^2)^{n/2}}$$

$\Longrightarrow$ in part (a)

$$\lambda(\mathbf{x}) = \frac{\max\limits_{\theta \in \Omega_H} L(\theta)}{\max\limits_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta)} = \left(\frac{s^2}{\sigma_0^2}\right)^{n/2} \times \exp\left\{-\frac{n}{2}\left[\left(\frac{s^2}{\sigma_0^2}\right) - 1\right]\right\} = g(t) \text{ say}$$

where $t = \frac{s^2}{\sigma_0^2}$ and $g(t) = t^{\nu} \exp\{-\nu(t-1)\}$, $\nu = n/2$.

LRT : Applications under Single Distributions

So the CR : $\lambda(\mathbf{x}) < c \iff \psi(t) < c_1$

where $\psi(t) = \log g(t) = \nu \log t - \nu(t-1)$

Now $\psi'(t) = 0 \implies \frac{\nu}{t} - \nu = 0$ or $t = 1$

Also $\psi''(t)|_{t=1} = -\frac{\nu}{1} < 0.$

So the S.O.C. $\implies \psi(t)$ is maximised (globally) at $t = 1$.

Hence the CR : $\psi(t) < c_1 \iff \{t < k_1 \text{ or } t > k_2\}$

LRT : Applications under Single Distributions

$$\Longleftrightarrow \left\{ \frac{nS^2}{\sigma_0^2} < \lambda_1 \text{ or } \frac{nS^2}{\sigma_0^2} > \lambda_2 \right\}$$

where $0 < \lambda_1 < \lambda_2 \ni$

$$P_H\left[\frac{nS^2}{\sigma_0^2} < \lambda_1\right] + P_H\left[\frac{nS^2}{\sigma_0^2} > \lambda_2\right] = \alpha(\text{size condition})$$

As $\frac{nS^2}{\sigma_0^2} \overset{H}{\sim} \chi^2_{n-1}$, assumption of equal tail probability $\Longrightarrow$

$\lambda_1 = \chi^2_{1-\frac{\alpha}{2};n-1}$, $\lambda_2 = \chi^2_{\frac{\alpha}{2};n-1}$

LRT : Applications under Single Distributions

$$(b)\,\Omega_H \cup \Omega_{K_2} = \{(\mu,\sigma) : \mu \in \mathcal{R}\,,\ \sigma_0 \le \sigma < \infty\}$$

$\Longrightarrow$ $\Omega_H \cup \Omega_{K_2}$: a restricted parameter space w.r.t. σ.

Here the ML estimate of σ^2 is

$$\begin{aligned} \sigma^2_{ml} &= s^2 \text{ if } \sigma_0^2 \le s^2 < \infty \\ &= \sigma_0^2 \text{ if } s^2 < \sigma_0^2 \end{aligned}$$

$\Longrightarrow$ accordingly

$$\begin{aligned} \max_{\theta \in \Omega_H \cup \Omega_{K_2}} L(\theta) &= \frac{\exp\{-n/2\}}{(2\pi s^2)^{n/2}} \text{ when } s^2 \ge \sigma_0^2 \\ &= \frac{\exp\{-n/2\}}{\left(2\pi\sigma_0^2\right)^{n/2}} \text{ when } s^2 < \sigma_0^2 \end{aligned}$$

LRT : Applications under Single Distributions

Hence the LR criterion will be

$$\begin{aligned}\lambda(\mathbf{x}) &= \left(\frac{s^2}{\sigma_0^2}\right)^{n/2} \times \exp\left\{-\frac{n}{2}\left[\left(\frac{s^2}{\sigma_0^2}\right)-1\right]\right\} \text{ if } s^2 \geq \sigma_0^2 \\ &= 1 \text{ otherwise}\end{aligned}$$

As $\lambda(\mathbf{x}) = 1 \implies$ the case of trivial acceptance of H.

$\implies$ the CR : $\lambda(\mathbf{x}) < c$ should be a subset of $\{s^2 \geq \sigma_0^2\} = \{t \geq 1\}$

Here also $\lambda(\mathbf{x}) < c \iff g(t) < c_1 \iff \psi(t) < c_2$ with $t \geq 1$ now.

LRT : Applications under Single Distributions

As $t = 1$ is the point of maximum for $g(t)$ or $\psi(t)$,

$\implies \psi(t) \downarrow t$ when $t \geq 1$.

Thus $\psi(t) < c_2 \iff t > k$.

$\implies$ the LRT critical region will be $\{t > k\}$ or $\left\{\frac{nS^2}{\sigma_0^2} > \lambda\right\}$

where $\lambda = \chi^2_{\alpha;n-1}$ due to size-α condition.

LRT : Applications under Single Distributions

Application 2: $X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential (mean $= \theta$).

To test $H : \theta = \theta_0$ versus (a) $K_1 : \theta \neq \theta_0$, (b) $K_2 : \theta < \theta_0$.

$\implies$ here $\Omega_H = \{\theta_0\}$, $\Omega_{K_1} = \{\theta : \theta > 0 \, , \, \theta \neq \theta_0\}$ and $\{\theta : 0 < \theta < \theta_0\}$.

The likelihood function : $L(\theta) = \frac{1}{\theta^n} \exp\left\{-\frac{\sum x_i}{\theta}\right\}$

$$\implies \max_{\theta \in \Omega_H} L(\theta) = \frac{1}{\theta_0^n} \exp\left\{-\frac{\sum x_i}{\theta_0}\right\} = \frac{1}{\theta_0^n} \exp\left\{-\frac{n\overline{x}}{\theta_0}\right\}.$$

LRT : Applications under Single Distributions

(a) $\Omega_H \cup \Omega_{K_1} = \{\theta : \theta > 0\} \longrightarrow$ unrestricted parameter space of θ.

$\implies$ the MLE of θ under $\Omega_H \cup \Omega_{K_1}$ is as usual $\overline{x}$.

$\implies \max\limits_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) = \frac{1}{\overline{x}^n} \exp\left\{-\frac{n\overline{x}}{\overline{x}}\right\} = \frac{1}{\overline{x}^n} \exp\{-n\}$.

As a result the LR statistic :

$$\lambda(\mathbf{x}) = \left(\frac{\overline{x}}{\theta_0}\right)^n \exp\left\{-n\left[\left(\frac{\overline{x}}{\theta_0}\right) - 1\right]\right\} = g(t)\,(\text{say})$$

where $t = \frac{\overline{x}}{\theta_0}$ and $g(t) = t^n \exp\{-n(t-1)\}$.

LRT : Applications under Single Distributions

So the CR :

$$\lambda(\mathbf{x}) < c \iff \psi(t) < c_1$$

where $\psi(t) = \ln g(t) = n \ln t - n(t-1)$.

Here also $\psi(t)$ gets maximised globally at $t = 1$.

Thus CR : $\psi(t) < c_1 \iff \{t < k_1\} \cup \{t > k_2\}$, or equivalently

$$\left\{ \frac{2\sum x_i}{\theta_0} < \lambda_1 \quad \text{or} \quad \frac{2\sum x_i}{\theta_0} > \lambda_2 \right\}$$

LRT : Applications under Single Distributions

As X_is are *iid* exponential(mean $= \theta$)

$$\implies \chi^2_H = \frac{2\sum X_i}{\theta_0} \overset{H}{\sim} \chi^2_{2n}.$$

$\implies$ the LRT CR : $\left\{\chi^2_H < \lambda_1 \ \text{ or } \ \chi^2_H > \lambda_2\right\}$ where $0 < \lambda_1 < \lambda_2$ are $\ni$

$P_H\left[\chi^2_H < \lambda_1\right] + P_H\left[\chi^2_H > \lambda_2\right] = \alpha$ (*size-α condition*)

Finally assume equal tail probability for simplicity and get

$\lambda_1 = \chi^2_{1-\frac{\alpha}{2};2n}$ and $\lambda_2 = \chi^2_{\frac{\alpha}{2};2n}$.

LRT : Applications under Single Distributions

(b) $\Omega_H \cup \Omega_{K_2} = \{\theta : 0 < \theta \leq \theta_0\} \longrightarrow$ restricted parameter space of θ.

Thus MLE of θ is

$$\begin{aligned}\hat{\theta} &= \overline{x} \quad , 0 < \overline{x} \leq \theta_0 \\ &= \theta_0 \quad , \overline{x} > \theta_0.\end{aligned}$$

Accordingly

$$\begin{aligned}\max_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) &= \frac{1}{\overline{x}^n} \exp\{-n\} \ , \ 0 < \overline{x} \leq \theta_0 \\ &= \frac{1}{\theta_0^n} \exp\left\{-\frac{n\overline{x}}{\theta_0}\right\} \ , \ \overline{x} > \theta_0\end{aligned}$$

$$\begin{aligned}\Longrightarrow \lambda(\mathbf{x}) &= \left(\frac{\overline{x}}{\theta_0}\right)^n \exp\left\{-n\left[\left(\frac{\overline{x}}{\theta_0}\right) - 1\right]\right\}, 0 < \overline{x} \leq \theta_0 \\ &= 1 \quad , \quad \overline{x} > \theta_0\end{aligned}$$

LRT : Applications under Single Distributions

But $\{\lambda(\mathbf{x}) = 1\}$ or $\{\overline{x} > \theta_0\}$ $\longrightarrow$ region of trivial acceptance of H.

$\implies$ the CR : $\{\lambda(\mathbf{x}) < c\} \subseteq \{\mathbf{x} : 0 < \overline{x} \leq \theta_0\}$ or $\{t \leq 1\}$ where $t = \frac{\overline{x}}{\theta_0}$.

$\implies$ as earlier, here also

$$\lambda(\mathbf{x}) < c \iff g(t) < c_1 \iff \psi(t) < c_2$$

Moreover now $\psi(t) \uparrow t \, \forall \, t \leq 1$.

So CR : $\{\psi(t) < c_2\} \iff \{t < k\} = \left\{ \frac{2\sum x_i}{\theta_0} \right\}$

where $\lambda = \chi^2_{1-\alpha;2n}$ due to size-α condition.

LRT : Applications under Single Distributions

In this module, the two applications, discussed so far, are based on the distributions belonging to OPEF. Now we will consider the distributions which are not members of OPEF.

Application 3. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ common pdf $f_\theta(x) = \frac{\theta}{x^2}$, $x \geq \theta$; $\theta > 0$. Suppose we are to test $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$.

Here $\Omega_H = \{\theta : 0 < \theta \leq \theta_0\}$, a restricted parameter space and $\Omega_H \cup \Omega_K = \{\theta : 0 < \theta < \infty\}$, the unrestricted parameter space.

The likelihood function of θ is $L(\theta) = \frac{\theta^n}{\left(\prod x_i\right)^2}$; $0 < \theta \leq x_{(1)} \leq \ldots \leq x_{(n)}$

$$
\begin{aligned}
\text{i.e. } L(\theta) &= \frac{\theta^n}{\left(\prod x_i\right)^2} \text{ ; } 0 < \theta \leq x_{(1)} \\
&= 0 \text{ , otherwise}
\end{aligned}
$$

LRT : Applications under Single Distributions

$\implies L(\theta) \uparrow \theta$ for $0 < \theta \leq x_{(1)}$.

Hence $L(\theta)$ attains its maximum at $\theta = x_{(1)} = \hat{\theta}$, say. Thus

$$\sup_{\Omega_H \cup \Omega_K} L(\theta) = \frac{\hat{\theta}^n}{\left(\prod x_i\right)^2}.$$

Also under Ω_H, ML estimate of θ is

$$\begin{aligned} \theta_{ml} &= \hat{\theta}\,,\ \hat{\theta} \leq \theta_0 \\ &= \theta_0\,,\ \hat{\theta} > \theta_0 \end{aligned}$$

$$\begin{aligned} \implies \sup_{\Omega_H} L(\theta) &= \frac{\hat{\theta}^n}{\left(\prod x_i\right)^2}\,,\ \hat{\theta} \leq \theta_0 \\ &= \frac{\theta_0^n}{\left(\prod x_i\right)^2}\,,\ \hat{\theta} > \theta_0 \end{aligned}$$

LRT : Applications under Single Distributions

So the LR criterion is

$$\lambda(\mathbf{x}) = 1 \;,\; \hat{\theta} \leq \theta_0$$
$$= \left(\frac{\theta_0}{\hat{\theta}}\right)^n < 1 ,\; \hat{\theta} > \theta_0$$

$\implies \lambda(\mathbf{x}) \downarrow \hat{\theta}.$

$\implies$ the LR test CR : $\lambda(\mathbf{x}) < \lambda \iff \hat{\theta} > c$, where c is such that

$$\sup_{\Omega_H} P_\theta[\hat{\theta} > c] = \alpha$$

where α : given level of significance. i.e.

$$\sup_{\theta \leq \theta_0} P_\theta[X_{(1)} > c] = \alpha$$

LRT : Applications under Single Distributions

Now $P_\theta[X_{(1)} > c] = (P_\theta[X_1 > c])^n = \left(\int_c^\infty \frac{\theta}{x^2} dx\right)^n = \left(\frac{\theta}{c}\right)^n \uparrow \theta.$

$$\implies \alpha = \sup_{\theta \le \theta_0} P_\theta[X_{(1)} > c] = \left(\frac{\theta_0}{c}\right)^n$$

i.e. $c = \frac{\theta_0}{\alpha^{1/n}}$ and hence

the CR under LRT is $\left\{X_{(1)} > \frac{\theta_0}{\alpha^{1/n}}\right\}$.

LRT : Applications under Single Distributions

Application 4. Suppose a single observation $X \sim$ Cauchy$(0, \sigma)$ distribution with pdf

$$f(x) = \frac{\sigma}{\pi \left\{\sigma^2 + x^2\right\}}\ ,\ -\infty < x < \infty\, ;\, \sigma > 0.$$

Based on this observation only suppose we are to test $H : \sigma \leq \sigma_0$ versus $K : \sigma > \sigma_0$.

Here $\Omega_H = \{\sigma : 0 < \sigma \leq \sigma_0\} \longrightarrow$ a restricted parameter space, and $\Omega_H \cup \Omega_K = \{\sigma : 0 < \sigma < \infty\} \longrightarrow$ unrestricted parameter space.

The likelihood function of σ is $L(\sigma) = \frac{\sigma}{\pi\{\sigma^2+x^2\}}$, $0 < \sigma < \infty$.

$\implies \ell(\theta) = \ln L(\theta) = \ln \sigma - \ln \pi - \ln \left\{\sigma^2 + x^2\right\}$

Hence one will get that under $\Omega_H \cup \Omega K$, $\sigma_{ml} = |x|$. (*Try yourself*)

LRT : Applications under Single Distributions

$$\implies \sup_{\Omega_H \cup \Omega K} L(\sigma) = \frac{1}{2\pi|x|} \text{ (putting } \sigma = |x| \text{ at } L(\sigma) \text{)}$$

Also under Ω_H

$$\begin{aligned}
\sigma_{ml} &= |x| \ , \ |x| \leq \sigma_0 \\
&= \sigma_0 \ , \ |x| > \sigma_0 \\
\implies \sup_{\Omega_H} L(\sigma) &= \frac{1}{2\pi|x|} \ , \ |x| \leq \sigma_0 \\
&= \frac{\sigma_0}{\pi\left\{\sigma_0^2 + |x|^2\right\}} \ , \ |x| > \sigma_0
\end{aligned}$$

So the LR criterion will be

$$\begin{aligned}
\lambda(x) &= 1 \ , \ |x| \leq \sigma_0 \\
&= \frac{2\sigma_0|x|}{\sigma_0^2 + |x|^2} < 1 \ , \ |x| > \sigma_0
\end{aligned}$$

LRT : Applications under Single Distributions

Thus $\lambda(x) \downarrow |x|$.

So the CR : $\lambda(x) < \lambda \iff |x| > c$, where c is $\ni$

$$\sup_{\sigma \leq \sigma_0} P_\sigma[|X| > c] = \alpha \,(\text{known significance level}) \ldots (*)$$

$$\begin{aligned} \text{Now} P_\sigma[|X| \leq c] &= 2\int_0^c \frac{\sigma}{\pi\left\{\sigma^2 + x^2\right\}} dx \\ &= \frac{2}{\pi} \tan^{-1}\left(\frac{c}{\sigma}\right) \end{aligned}$$

$\implies P_\sigma[|X| > c] = 1 - \frac{2}{\pi} \tan^{-1}\left(\frac{c}{\sigma}\right) \uparrow \sigma.$

LRT : Applications under Single Distributions

$$\implies \sup_{\sigma \leq \sigma_0} P_\sigma[|X| > c] = P_{\sigma_0}[|X| > c] = 1 - \frac{2}{\pi}\tan^{-1}\left(\frac{c}{\sigma_0}\right).$$

hence from $(*)$ we get $1 - \frac{2}{\pi}\tan^{-1}\left(\frac{c}{\sigma_0}\right) = \alpha$

Or $\tan^{-1}\left(\frac{c}{\sigma_0}\right) = (1-\alpha)\frac{\pi}{2}$

i.e. $c = \sigma_0 \tan[(1-\alpha)\frac{\pi}{2}]$.

So the CR for LRT is $\left\{|X| > \sigma_0 \tan[(1-\alpha)\frac{\pi}{2}]\right\}$.

LRT : Applications under Single Distributions

TRY YOURSELF!

11\. 1. Suppose a single observation is drawn from a Cauchy $(0, \sigma)$ population. Discuss an LRT for testing $H : \sigma = \sigma_0$ versus $K : \sigma \neq \sigma_0$. If the observation is - 0. 2978 and $\sigma_0 = 1$, what will be your dicision about acceptance or rejection of H?

11\. 2. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ common pdf $f_\theta(x) = \frac{\theta}{x^2}$, $x \geq \theta$; $\theta > 0$. Suppose we are to test $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$. Discuss an LRT for testing these hypotheses.
(*Readers themselves can solve it directly if they go through* **Application 3.** *of this Module*)

LRT : Applications under Single Distributions

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 11 . . .

M11. 1. The basic set up of this problem is same as that in **Application 4.** So we will use the expressions and findings of that application directly in this problem as and when necessary.

Here $\Omega_H = \{\sigma = \sigma_0\}$, a singleton set, and $\Omega_H \cup \Omega_K = \{\sigma : 0 < \sigma < \infty\}$, unrestricted parameter space of σ.

$$\implies \sup_{\Omega_H} L(\sigma) = \frac{\sigma_0}{\pi\left\{\sigma_0^2 + x^2\right\}}$$

and

$$\sup_{\Omega_H \cup \Omega_K} L(\sigma) = \frac{1}{2\pi|x|}$$

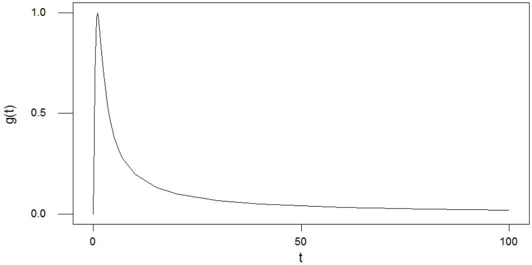
(since the MLE of σ under $\Omega_H \cup \Omega_K$ is $|x|$, using it in $L(\sigma)$ we get the supremum of the likelihood as above)

LRT : Applications under Single Distributions

So the LR criterion: $\lambda(x) = \frac{2\sigma_0|x|}{\sigma_0^2+|x|^2} = g(t)$, say where

$$g(t) = \frac{2\sigma_0 t}{\sigma_0^2 + t^2}\ , \text{ and } t = |x|.$$

Now check that $g(t)$ is globally maximised at $t = \sigma_0$ and t versus $g(t)$ plot appears as follows:



LRT : Applications under Single Distributions

Hence the LR critical region $\lambda(x)\,(= g(t)) < \lambda \iff \{T < c_1 \text{ or } T > c_2\}$, where $T = |X|$ and $c_1 < c_2$ are such that

$g(c_1) = g(c_2)$ and $P_{\sigma_0}[T < c_1] + P_{\sigma_0}[T > c_2] = \alpha$ (given significance level)

Now $g(c_1) = g(c_2) \implies \frac{c_1}{\sigma_0^2 + c_1^2} = \frac{c_2}{\sigma_0^2 + c_2^2}$

which on simplification finally gives $c_1 c_2 = \sigma_0^2$ (as $c_1 \neq c_2$); i.e. $c_2 = \frac{\sigma_0^2}{c_1}$.

$\implies P_{\sigma_0}[T < c_1] + P_{\sigma_0}[T > \frac{\sigma_0^2}{c_1}] = \alpha.$

$$\implies \frac{2}{\pi}\tan^{-1}\frac{c_1}{\sigma_0} + 1 - \frac{2}{\pi}\tan^{-1}\frac{\sigma_0}{c_1} = \alpha$$

(as we already know that $P_{\sigma_0}[T \leq a] = \frac{2}{\pi}\tan^{-1}\frac{a}{\sigma_0}$)

LRT : Applications under Single Distributions

On simplification the relation becomes

$$\frac{2}{\pi}\tan^{-1}\frac{\sigma_0^2 - c_1^2}{2c_1\sigma_0} = 1-\alpha$$

$$\text{i.e. } \frac{\sigma_0^2 - c_1^2}{2c_1\sigma_0} = \tan\left[(1-\alpha)\frac{\pi}{2}\right]$$

$$\implies \sigma_0^2 - c_1^2 = 2\sigma_0\gamma c_1$$

where $\gamma = \tan\left[(1-\alpha)\frac{\pi}{2}\right]$ is known for given α. So finally we have

$$c_1^2 + 2\sigma_0\gamma c_1 - \sigma_0^2 = 0 \quad \ldots \quad \ldots \quad (**)$$

The solution of $(**)$ considered feasible to this problem, gives the cut-off points of the test.

LRT : Applications under Single Distributions

For the second part of this problem, we solve $(**)$ for c_1 and hence for c_2 also, using R-programing under the choice $\sigma_0 = 1$ and $\alpha = 0.05, 0.1$.
Finally in the light of the given value, $x = -2.9748$ we will decide about acceptance or rejection of H for both the levels.

LRT : Applications under Single Distributions

Computation using R : (to find 'c_1' and 'c_2')

R Code and Output :

```
> iowa=function(sigma,alpha)
{
gamma=tan((1-alpha)*(0.5*pi))
gamma
a=1
b=2*gamma*sigma
c=-(sigma*sigma)
del=(b*b)-(4*a*c)
c11=((-b)+sqrt(del))/(2*a)
c12=((-b)-sqrt(del))/(2*a)
c21=(sigma*sigma)/c11
c22=(sigma*sigma)/c12
```

LRT : Applications under Single Distributions

R Code and Output (continued) :

```
if(c11>0)
{
cat(c11,"\n",c21)
cat("\n\n")
}
else if(c12>0)
{
cat(c12,"\n",c22)
cat("\n\n")
}
}
> # values of 'c1' and 'c2' for different values of 'alpha'
> iowa(1,0.05)
0.03929011
 25.4517
> iowa(1,0.1)
0.07870171
 12.7062
```

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Module 12

Saurav De

Department of Statistics
Presidency University

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Application 1: $X_1, \ldots, X_m \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

Suppose we want to test $H : \sigma_1 = \sigma_2$ versus
$(a)\ K_1 : \sigma_1 \neq \sigma_2$, $(b)\ K_2 : \sigma_2 > \sigma_1$

Here $\theta \equiv (\mu_1, \mu_2, \sigma_1, \sigma_2)$.

$$\Omega_H = \{\theta : \mu_1, \mu_2 \in \mathcal{R}\, , \, \sigma_2 = \sigma_1 = \sigma, \sigma > 0\}\, ,$$

$\Omega_H \cup \Omega_{K_1} = \{\theta : \mu_1, \mu_2 \in \mathcal{R}\, , \, \sigma_1, \sigma_2 > 0\} \longrightarrow$ fully unrestricted parameter space, and

$\Omega_H \cup \Omega_{K_2} = \{\theta : \mu_1, \mu_2 \in \mathcal{R}\, , \, \sigma_2 \geq \sigma_1 > 0\} \longrightarrow$ partially restricted parameter space w.r.t. σ_1, σ_2.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

The likelihood under $\Omega_H \cup \Omega_{K_1}$

$$L(\theta) = C(\sigma_1^2, \sigma_2^2) \exp\left\{-\frac{1}{2}\left[\frac{\sum(x_i-\mu_1)^2}{\sigma_1^2} + \frac{\sum(y_j-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

$$= C(\sigma_1^2, \sigma_2^2) \exp\left\{-\frac{1}{2}\left[\frac{ms_1^2 + m(\overline{x}-\mu_1)^2}{\sigma_1^2} + \frac{ns_2^2 + n(\overline{y}-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

where $C(\sigma_1^2, \sigma_2^2) = \frac{1}{(\sqrt{2\pi})^{m+n}(\sigma_1^2)^{m/2}(\sigma_2^2)^{n/2}}$, s_1^2 and s_2^2 are the sample variances of X and Y with divisors m and n respectively.

(a) Under $\Omega_H \cup \Omega_{K_1}$ MLE of θ is $(\overline{x}, \overline{y}, s_1^2, s_2^2)$.

$$\implies \max_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) = \frac{1}{(\sqrt{2\pi})^{m+n}(s_1^2)^{m/2}(s_2^2)^{n/2}} \exp\left\{-\frac{m+n}{2}\right\}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Also under Ω_H

$$L(\theta) = \frac{1}{(\sqrt{2\pi})^{m+n}(\sigma^2)^{\frac{m+n}{2}}}$$
$$\exp\left\{-\frac{1}{2\sigma^2}\left[ms_1^2 + m(\overline{x}-\mu_1)^2 + ns_2^2 + n(\overline{y}-\mu_2)^2\right]\right\}$$

$\implies$ MLE of σ^2 will be $\hat{\sigma}^2 = \frac{ms_1^2+ns_2^2}{m+n}$ = pooled sample variance.

$$\implies \max_{\theta\in\Omega_H} L(\theta) = \frac{1}{(\sqrt{2\pi})^{m+n}(\hat{\sigma}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\}.$$

So the LR criterion: $\implies \lambda(\mathbf{x}) = \frac{(s_1^2)^{m/2}(s_2^2)^{n/2}}{(\hat{\sigma}^2)^{\frac{m+n}{2}}} = \frac{t^{n/2}(m+n)^{\frac{m+n}{2}}}{(nt+m)^{\frac{m+n}{2}}}$

where $t = \frac{s_2^2}{s_1^2}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Now

$$\lambda(\mathbf{x}) < c \iff g(t) = \frac{t^{n/2}}{(nt+m)^{\frac{m+n}{2}}} < c_1 \iff \psi(t) < c_2$$

where $\psi(t) = \ln g(t) = \frac{n}{2}\ln t - \frac{m+n}{2}\ln(m+nt)$.

$\psi'(t) = 0 \implies \frac{n}{2t} - \frac{(m+n)}{2}\frac{n}{(m+nt)} = 0 \iff t = 1$ (on simplification)

Also here $\psi''(t)|_{t=1} < 0$

$\implies \psi(t)$ has global maximum at $t = 1$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Thus $\psi(t) < c_2 \iff \{t < k_1 \text{ or } t > k_2\}$, $k_1 < k_2$.

$$\iff \{F_H < \lambda_1 \text{ or } F_H > \lambda_2\} \text{ , } \lambda_1 < \lambda_2.$$

where $F_H = \frac{\frac{ns_2^2}{n-1}}{\frac{ms_1^2}{m-1}} \overset{H}{\sim} F_{n-1,m-1}$ and λ_1, λ_2 are such that

$$P_H\left[F_H < \lambda_1\right] + P_H\left[F_H > \lambda_2\right] = \alpha \text{ (size} - \alpha \text{ condition)} \ldots (*)$$

and $\psi(k_1) = \psi(k_2)$ or $g(k_1) = g(k_2)$ i.e.

$$\frac{k_1^{n/2}}{(nk_1+m)^{\frac{m+n}{2}}} = \frac{k_2^{n/2}}{(nk_2+m)^{\frac{m+n}{2}}} \ldots \ldots (**)$$

Equal tail probability assumption doesn't need $(**)$ and simplifies the solution of $(*)$ with $\lambda_1 = F_{1-\frac{\alpha}{2};n-1,m-1}$ and $\lambda_2 = F_{\frac{\alpha}{2};n-1,m-1}$, though it no longer remains an LRT then.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

(b) Under $\Omega_H \cup \Omega_{K_2} = \left\{\theta : \mu_1, \mu_2 \in \mathcal{R}, \frac{\sigma_2^2}{\sigma_1^2} \geq 1\right\}$

$$\begin{aligned}
\text{MLE of } \frac{\sigma_2^2}{\sigma_1^2} &= \frac{s_2^2}{s_1^2}, \ \frac{s_2^2}{s_1^2} \geq 1 \\
&= 1, \ \frac{s_2^2}{s_1^2} < 1
\end{aligned}$$

As $L(\theta)$ under $\Omega_H \cup \Omega_{K_2} = L(\theta)$ under $\Omega_H \cup \Omega_{K_1}$, this implies

$$\begin{aligned}
\max_{\theta \in \Omega_H \cup \Omega_{K_2}} L(\theta) &= \frac{1}{(\sqrt{2\pi})^{m+n}(s_1^2)^{m/2}(s_2^2)^{n/2}} \exp\left\{-\frac{m+n}{2}\right\}, \ \frac{s_2^2}{s_1^2} \geq 1 \\
&= \frac{1}{(\sqrt{2\pi})^{m+n}(\hat{\sigma}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\}, \ \frac{s_2^2}{s_1^2} < 1
\end{aligned}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Accordingly

$$\lambda(\mathbf{x}) = \frac{(s_1^2)^{m/2}(s_2^2)^{n/2}}{(\hat{\sigma}^2)^{\frac{m+n}{2}}} \;,\; \frac{s_2^2}{s_1^2} \geq 1$$

$$= 1 \;,\; \frac{s_2^2}{s_1^2} < 1 \text{ (case of trivial acceptance of } H\text{)}$$

Thus the CR : $\lambda(\mathbf{x}) < c \subseteq \{t \geq 1\}$, $t = \frac{s_2^2}{s_1^2}$.

Now $\lambda(\mathbf{x}) < c \iff g(t) = \frac{t^{n/2}}{(nt+m)^{\frac{m+n}{2}}} < c_1 \iff \psi(t) = \ln g(t) < c_2$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

But $\psi(t)$ is maximum at $t = 1$ and $\psi(t) \downarrow t$ for $t \geq 1$.

Thus $\psi(t) < c_2 \iff \{t > k\} \iff \left\{ \frac{\frac{ns_2^2}{n-1}}{\frac{ms_1^2}{m-1}} > \lambda \right\}$

i.e. the LRT CR : $\{F_H > \lambda\}$, $F_H = \frac{\frac{ns_2^2}{n-1}}{\frac{ms_1^2}{m-1}} \overset{H}{\sim} F_{n-1,m-1}$.

The size-α condition gives $\lambda = F_{\alpha;n-1,m-1}$.

Note : H is accepted $\implies$ homoscedasticity i.e. equal variance holds.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Second Part Now, given homoscedasticity, consider the following hypotheses to test:

$H : \mu_2 - \mu_1 = 0$ versus $(a)\, K_1 : \mu_2 - \mu_1 \neq 0$, $(b)\, K_2 : \mu_2 - \mu_1 > 0$

$\Longrightarrow$ here the parameter $\theta = (\mu_1, \mu_2, \sigma^2)$, σ^2 : the common variance.

$\Omega_H = \{\theta : \mu_1, \mu_2 \in \mathcal{R}, \mu_2 = \mu_1, \sigma > 0\}$

$\Omega_H \cup \Omega_{K_1} = \{\theta : \mu_1, \mu_2 \in \mathcal{R}, \sigma > 0\}$ and

$\Omega_H \cup \Omega_{K_2} = \{\theta : -\infty < \mu_1 \leq \mu_2 < \infty, \sigma > 0\} \longrightarrow$ restricted parameter space w.r.t. (μ_1, μ_2).

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

The likelihood under $\Omega_H \cup \Omega_{K_1}$:

$$L(\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[ms_1^2 + m(\overline{x} - \mu_1)^2 + ns_2^2 + n(\overline{y} - \mu_2)^2\right]\right\}$$

We get MLE of (μ_1, μ_2, σ^2) as $(\overline{x}, \overline{y}, \hat{\sigma}^2)$, $\hat{\sigma}^2 = \frac{ms_1^2 + ns_2^2}{m+n}$.

$$\implies \max_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Under Ω_H : let $\mu_1 = \mu_2 = \mu$

$$\implies L(\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[ms_1^2 + m(\overline{x}-\mu)^2 + ns_2^2 + n(\overline{y}-\mu)^2\right]\right\}$$

Under Ω_H the MLE of μ is $\hat{\mu} = \frac{m\overline{x}+n\overline{y}}{m+n}$ and MLE of σ^2 is

$$\hat{\hat{\sigma}}^2 = \frac{ms_1^2 + ns_2^2 + md_1^2 + nd_2^2}{m+n}$$

where $d_1 = \overline{x} - \hat{\mu}$, $d_2 = \overline{y} - \hat{\mu}$.

$$\implies \max_{\theta \in \Omega_H} L(\theta) = \frac{1}{(2\pi\hat{\hat{\sigma}}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

$$\implies \lambda(\mathbf{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2}\right)^{\frac{m+n}{2}}$$

$$\text{The CR} : \lambda(\mathbf{x}) < c \iff \frac{\frac{ms_1^2+ns_2^2}{m+n}}{\frac{ms_1^2+ns_2^2+md_1^2+nd_2^2}{m+n}} < c_1$$

$$\iff 1 + \frac{md_1^2 + nd_2^2}{ms_1^2 + ns_2^2} > c_2$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

On simplification it becomes

$$\frac{(\overline{y}-\overline{x})^2}{s^2\left(\frac{1}{m}+\frac{1}{n}\right)} > c_3 \ , \ s^2 = \frac{ms_1^2+ns_2^2}{m+n}$$

$$\text{Or } \left| \frac{\overline{y}-\overline{x}}{\sqrt{s^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \right| > \lambda$$

where

$$t_H = \frac{\overline{Y}-\overline{X}}{\sqrt{S^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \overset{H}{\sim} t_{m+n-2}$$

The critical point $\lambda = t_{\frac{\alpha}{2};m+n-2}$ at level α.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

(b) Under $\Omega_H \cup \Omega_{K_2}$ where $0 \leq \mu_2 - \mu_1 < \infty$

$$\begin{aligned} \text{MLE of } \mu_2 - \mu_1 &= \bar{y} - \bar{x} \,,\ \bar{y} - \bar{x} \geq 0 \\ &= 0 \,,\ \text{otherwise} \end{aligned}$$

Thus

$$\begin{aligned} \max_{\theta \in \Omega_H \cup \Omega_{K_2}} L(\theta) &= \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\} \,,\ \bar{y} - \bar{x} \geq 0 \\ &= \frac{1}{(2\pi\hat{\hat{\sigma}}^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{m+n}{2}\right\} \,,\ \bar{y} - \bar{x} < 0 \end{aligned}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Accordingly

$$\lambda(\mathbf{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2}\right)^{\frac{m+n}{2}}, \ \overline{y} - \overline{x} \geq 0$$
$$= 1 \ , \ \text{(case when H is trivially accepted)}$$

$\implies$ CR : $\lambda(\mathbf{x}) < c \subseteq \{\overline{y} - \overline{x} \geq 0\}$

Now $\lambda(\mathbf{x}) < c \iff \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} < c_1 \implies$ on simplification

$$\frac{\overline{y} - \overline{x}}{\sqrt{s^2\left(\frac{1}{m} + \frac{1}{n}\right)}} > \lambda \ [\text{as here } \overline{y} - \overline{x} \geq 0]$$

Size-α condition $\implies \lambda = t_{\alpha;m+n-2}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Application 2: Let $X_1, \ldots, X_m \overset{iid}{\sim}$ common pdf $\frac{\theta_1}{x^2}$, $x \geq \theta_1 > 0$ and $Y_1, \ldots, Y_n \overset{iid}{\sim}$ common pdf $\frac{\theta_2}{y^2}$, $y \geq \theta_2 > 0$ independently.

Suppose we are to test $H : \theta_1 = \theta_2$ against $K : \theta_1 \neq \theta_2$
Here $\Omega_H = \{\theta : \theta_1 = \theta_2 = \theta \,;\, 0 < \theta < \infty ; \}$ and
$\Omega_H \cup \Omega_K = \{(\theta_1, \theta_2) : 0 < \theta_i < \infty \,,\, i = 1, 2\}$.

Under $\Omega_H \cup \Omega_K$ the likelihood function is

$$L(\theta_1, \theta_2) = \frac{\theta_1^m \theta_2^n}{\left(\prod_{i=1}^{m} x_i \prod_{j=1}^{n} y_j \right)^2} \,;\, \theta_1 \leq x_{(1)} \,,\, \theta_2 \leq y_{(1)}.$$

Obviously $L(\theta_1, \theta_2) \uparrow \theta_i \,,\, i = 1, 2.$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Hence MLE of (θ_1, θ_2) is $(X_{(1)}, Y_{(1)})$. $\implies$

$$\sup_{\Omega_H \cup \Omega_K} L(\theta_1, \theta_2) = \frac{x_{(1)}^m y_{(1)}^n}{\left(\prod_{i=1}^{m} x_i \prod_{j=1}^{n} y_j\right)^2}.$$

Similarly under Ω_H the likelihood function is

$$L(\theta) = \frac{\theta^{m+n}}{\left(\prod_{i=1}^{m} x_i \prod_{j=1}^{n} y_j\right)^2} ; \theta \leq x_{(1)} , \theta \leq y_{(1)} \text{ i.e. } \theta \leq \min\left\{x_{(1)}, y_{(1)}\right\}.$$

Obviously $L(\theta) \uparrow \theta$. And hence MLE of θ is $\min\left\{X_{(1)}, Y_{(1)}\right\}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

$$\implies \sup_{\Omega_H} L(\theta) = \frac{\left(\min\{x_{(1)}, y_{(1)}\}\right)^{m+n}}{\left(\prod_{i=1}^{m} x_i \prod_{j=1}^{n} y_j\right)^2}.$$

So the LR criterion is

$$\lambda(\mathbf{x}, \mathbf{y}) = \frac{\sup_{\Omega_H} L(\theta)}{\sup_{\Omega_H \cup \Omega_K} L(\theta_1, \theta_2)} = \frac{\left(\min\left\{x_{(1)}, y_{(1)}\right\}\right)^{m+n}}{x_{(1)}^m y_{(1)}^n}.$$

Now
$$\frac{\left(\min\left\{x_{(1)}, y_{(1)}\right\}\right)^{m+n}}{x_{(1)}^m y_{(1)}^n} = \left(\frac{x_{(1)}}{y_{(1)}}\right)^n, \text{ if } \min\left\{x_{(1)}, y_{(1)}\right\} = x_{(1)}$$
$$= \left(\frac{y_{(1)}}{x_{(1)}}\right)^m, \text{ if } \min\left\{x_{(1)}, y_{(1)}\right\} = y_{(1)}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

So the critical region

$$\{\lambda(\mathbf{x},\mathbf{y}) < \lambda\} = \left\{\left(\frac{x_{(1)}}{y_{(1)}}\right)^n < \lambda \text{ Or } \left(\frac{y_{(1)}}{x_{(1)}}\right)^m < \lambda\right\} ;\, 0 < \lambda < 1.$$

$$i.e. = \left\{\frac{x_{(1)}}{y_{(1)}} < \lambda^{1/n} \text{ Or } \frac{x_{(1)}}{y_{(1)}} > \frac{1}{\lambda^{1/m}}\right\},$$

where λ is $\ni$ $P_H[\frac{X_{(1)}}{Y_{(1)}} < \lambda^{1/n}] + P_H[\frac{X_{(1)}}{Y_{(1)}} > \frac{1}{\lambda^{1/m}}] = \alpha$ (the given significance level).

Now under H the pdf of $\frac{X_{(1)}}{Y_{(1)}}$ at u will be

$$\begin{aligned} g(u) &= \frac{mn}{m+n}u^{n-1},\, 0 < u < 1 \\ &= \frac{mn}{m+n}\frac{1}{u^{m+1}},\, u \geq 1 \;\; (\textit{left to the readers as an excercise}) \end{aligned}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

So $P_H[\frac{X_{(1)}}{Y_{(1)}} < \lambda^{1/n}] = \int_0^{\lambda^{1/n}} \frac{mn}{m+n} u^{n-1} du = \frac{m\lambda}{m+n}$

and $P_H[\frac{X_{(1)}}{Y_{(1)}} > \frac{1}{\lambda^{1/m}}] = \int_{\frac{1}{\lambda^{1/m}}}^{\infty} \frac{mn}{m+n} \frac{1}{u^{m+1}} du = \frac{n\lambda}{m+n}$.

Thus we have from size condition that $\frac{m\lambda}{m+n} + \frac{n\lambda}{m+n} = \alpha$ i.e. $\lambda = \alpha$.

$\Longrightarrow$ the size-α CR of the LRT is $\left\{ \frac{X_{(1)}}{Y_{(1)}} < \alpha^{1/n} \text{ Or } \frac{X_{(1)}}{Y_{(1)}} > \frac{1}{\alpha^{1/m}} \right\}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Application 3 : Application on bivariate normal population :

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$$

Parameter $\theta \equiv (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

Null hypothesis $H : \rho = 0$ versus $(a)\, \rho \neq 0$ and $(b)\, \rho > 0$ for testing.

$\Longrightarrow \Omega_H = \{\theta : \mu_1, \mu_2 \in \mathcal{R}, \sigma_1, \sigma_2 > 0, \rho = 0\}$

$\Omega_H \cup \Omega_{K_1} = \{\theta : \mu_1, \mu_2 \in \mathcal{R}, \sigma_1, \sigma_2 > 0, |\rho| < 1\} \longrightarrow$ unrestricted parameter space.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Under $\Omega_H \cup \Omega_{K_1}$ the likelihood

$$L(\theta) = \frac{1}{(2\pi)^n(\sigma_1^2\sigma_2^2(1-\rho^2))^{n/2}} \exp\left\{-\frac{1}{2}Q(\theta)\right\}$$

where $Q(\theta) = \frac{1}{(1-\rho^2)}\left[\frac{\sum(x_i-\mu_1)^2}{\sigma_1^2} + \frac{\sum(y_i-\mu_2)^2}{\sigma_2^2} - \frac{2\rho\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right]$

$$= \frac{n}{(1-\rho^2)}\left[\frac{s_1^2+(\overline{x}-\mu_1)^2}{\sigma_1^2} + \frac{s_2^2+(\overline{y}-\mu_2)^2}{\sigma_2^2} - \frac{2\rho\{r\, s_1 s_2+(\overline{x}-\mu_1)(\overline{y}-\mu_2)\}}{\sigma_1\sigma_2}\right]$$

where r = sample correlation coefficient on (X, Y).

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Under unrestricted case, we know from Module 5 that, MLE of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ is $(\overline{x}, \overline{y}, s_1^2, s_2^2, r)$.

$$\implies \max_{\theta \in \Omega_H \cup \Omega_{K_1}} L(\theta) = \frac{1}{(2\pi)^n (s_1^2 s_2^2 (1 - r^2))^{n/2}} e^{-n}$$

Also under $\rho = 0, X$ and Y are independently distributed $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ variables. $\implies$ MLE of $(\mu_1, \sigma_1^2) = (\overline{x}, s_1^2)$ and MLE of $(\mu_2, \sigma_2^2) = (\overline{y}, s_2^2)$.

$$\implies \max_{\theta \in \Omega_H} L(\theta) = \frac{1}{(2\pi)^n (s_1^2 s_2^2)^{n/2}} \exp\left\{-\left(\frac{n}{2} + \frac{n}{2}\right)\right\}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

$$\implies \lambda(\mathbf{x}) = \frac{\max\limits_{\theta\in\Omega_H} L(\theta)}{\max\limits_{\theta\in\Omega_H\cup\Omega_{K_1}} L(\theta)} = (1-r^2)^{n/2} \downarrow |r|$$

Now $\lambda(\mathbf{x}) < c \iff |r| > c_1$ and $\sqrt{1-r^2} < c_2$.

$$\iff \frac{\sqrt{n-2}\,|r|}{\sqrt{1-r^2}} > \lambda \text{ i.e. } \left|\frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}}\right| > \lambda.$$

From sampling distribution we know

$$t_H = \frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}} \overset{H}{\sim} t_{n-2}$$

$\implies \lambda = t_{\frac{\alpha}{2};n-2}$ at level of significance α.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

(b) Under $\Omega_H \cup \Omega_{K_2} : \rho \geq 0 \longrightarrow$ restricted parameter space of ρ,

$$\begin{aligned} \text{MLE of } \rho &= r \ , \ r \geq 0 \\ &= 0 \ , \ r < 0 \end{aligned}$$

$$\begin{aligned} \Longrightarrow \max_{\theta \in \Omega_H \cup \Omega_{K_2}} L(\theta) &= \frac{1}{(2\pi)^n (s_1^2 s_2^2 (1-r^2))^{n/2}} e^{-n} \ , \ r \geq 0 \\ &= \frac{1}{(2\pi)^n (s_1^2 s_2^2)^{n/2}} e^{-n} \ , \ r < 0 \end{aligned}$$

$$\begin{aligned} \Longrightarrow \lambda(\mathbf{x}) &= (1-r^2)^{n/2} \ , \ r \geq 0 \\ &= 1 \ , \ r < 0 \text{ (case of trivial acceptance of H)} \end{aligned}$$

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

$\implies$ CR : $\lambda(\mathbf{x}) < c \subseteq \{r \geq 0\}$

Also $\lambda(\mathbf{x}) < c \iff 1 - r^2 < c_1 \iff r^2 > c_2$

$\iff r > c_3$ (as $r \geq 0$).

So the LRT CR : $\lambda(\mathbf{x}) < c \iff \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} > \lambda$.

Level of significance $= \alpha \implies \lambda = t_{\alpha;n-2}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Application 4 :

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$$

Parameter $\theta \equiv (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

Let the null hypothesis $H : \frac{\sigma_1^2}{\sigma_2^2} = \zeta_0^2$(known) versus $(a)\, K_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq \zeta_0^2$ and $(b)\, K_2 : \frac{\sigma_1^2}{\sigma_2^2} > \zeta_0^2$ for testing.

Let us make linear transformations like $U_i = X_i + \zeta_0 Y_i$ and $V_i = X_i - \zeta_0 Y_i$, $i = 1, 2, \ldots, n$.
$\implies (U_1, V_1), \ldots, (U_n, V_n)$ also follow *iid* BVN with population corr ρ_{uv}, say.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Obviously $Cov(U,V) = \sigma_1^2 - \zeta_0^2\sigma_2^2$. $\implies Cov(U,V) = 0$ under H. As a result $\rho_{uv} = 0$ under H. Hence in terms of the tranformed variables (U_i, V_i)s our testing problem boils down to testing

$$H : \rho_{uv} = 0 \text{ versus } (a)\, K_1 : \rho_{uv} \neq 0 \;(b)\, K_2 : \rho_{uv} > 0$$

Now proceeding exactly in the same way as the earlier application, the size-α LR critical region for H versus K_1 is

$$|t_H| > t_{\frac{\alpha}{2};n-2}$$

and the same for H versus K_2 is $t_H > t_{\alpha;n-2}$, where $t_H = \frac{\sqrt{n-2}r_{uv}}{\sqrt{1-r_{uv}^2}}$ following t_{n-2} distribution under H and r_{uv} : sample correlation coefficient on (U,V) based on n observations.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Application 5 : Let $X \sim N(\mu_1, 1)$ and $Y \sim$ Cauchy $(\mu_2, 1)$ independently. To test $H : \mu_1 = \mu_2 = 0$ versus K : not H. Here the likelihood is

$$L(\mu_1, \mu_2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu_1)^2\right\} \frac{1}{\pi\left\{1+(y-\mu_2)^2\right\}}$$

and hence the loglikelihood is

$$\ell(\mu_1, \mu_2) = \text{Const } -\frac{1}{2}(x-\mu_1)^2 - \ln\left\{1+(y-\mu_2)^2\right\}.$$

Now it would not be difficult to show that ML estimate of (μ_1, μ_2) is (x, y) which maximises L.

$$\implies \sup L(\mu_1, \mu_2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\pi}.$$

Also $\sup_H L(\mu_1, \mu_2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} . \frac{1}{\pi(1+y^2)}$.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

$\Longrightarrow$ the LR criterion is $\lambda(x,y) = \frac{\exp\{-\frac{1}{2}x^2\}}{(1+y^2)}$.

$\Longrightarrow$ the CR is $\lambda(x,y) < \lambda$ where λ is $\ni$

$$P_H[\lambda(X,Y) < \lambda] = \alpha \text{ (given significance level)}$$

$$i.e.\ P_H\left[\frac{\exp\left\{-\frac{1}{2}X^2\right\}}{(1+Y^2)} < \lambda\right] = \alpha \ldots\ldots\ (*)$$

where $X \overset{H}{\sim} \text{N}\ (0\ ,\ 1)$ and $Y \overset{H}{\sim} \text{Cauchy}\ (0\ ,\ 1)$ independently.

Note. Since here no explicit solution for λ is possible from the size condition $(*)$ hence using simulation technique through an R-program we can get the empirical solution for λ w.r.t. some given value of $\alpha = 0.05$ or 0.1 etc.

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

Computation using R : (to find suitable λ for $\alpha = 0.05$)

R Code and Output :

```
> library(varhandle)
> lambda=NULL
> for(j in 1:1000)
{
x=rnorm(10000,0,1)
y=rcauchy(10000,0,1)
W=exp(-(x*x)/2)/(1 + (y*y))
W1=sort(W)
mat=as.data.frame(table(W1))
CF=NULL
sum=0
```

LRT : Applications under Two Comparable Single Distributions and Bivariate Distribution

R Code and Output (continued) :

```
for(i in 1:length(mat[,1]))
{
sum=sum+mat[,2][i]
CF[i]=sum
}
mat1=cbind(mat,CF)
lambda[j]=unfactor(mat1[,1][which(mat1[,3]==500)])
x=NULL
y=NULL
W=NULL
}
> # usable lambda/cut-off value
> mean(lambda)
[1] 0.003076692
```

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Module 13

Saurav De

Department of Statistics
Presidency University

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Application 1.

- Consider k (> 2) independent normal populations $N(\mu_1, \sigma_1^2), \dots, N(\mu_k, \sigma_k^2)$.
- Let $\{x_{i1}, \dots, x_{in_i}\}$ be a random sample of size n_i drawn from $N(\mu_i, \sigma_i^2), \quad i = 1, \dots, k$.
- $\sum_{i=1}^{k} n_i = n$ is the total sample size.
- Suppose we want to test $H : \mu_1 = \cdots = \mu_k = \mu$ (say) versus alternative $K : \text{not } H$.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

- To avoid the Behren-Fisher's type problems, we have to assume $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ (say) (assumption of homoscedasticity).
- Under homoscedasticity assumption, $\theta = (\mu_1, \ldots, \mu_k, \sigma)$ and $\Omega_H = \{(\mu_1, \ldots, \mu_k, \sigma^2) : \mu_i = \mu, \; -\infty < \mu < \infty; \sigma > 0\}$ and $\Omega_H \cup \Omega_K = \{(\mu_1, \ldots, \mu_k, \sigma^2) : -\infty < \mu_i < \infty, i = 1, \ldots, k; \sigma > 0\}$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

- Now the distribution of θ under $\Omega_H \cup \Omega_K$ is :

$$L(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}.$$

- With the straightforward or routine derivation, we get the MLE of μ_i as : $\hat{\mu}_i = \overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ and the MLE of σ^2 as : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2 = \frac{SSW}{n}$.
- Similarly under Ω_H, $L(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu)^2}$.
- It can be shown here, i.e., under Ω_H, the MLE of μ is : $\hat{\hat{\mu}} = \overline{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$ and the MLE of σ^2 is : $\hat{\hat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = \frac{TotalSS}{n} = \frac{TSS}{n}$.

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

$$\text{Hence, } \max_{\theta \in \Omega_H \cup \Omega_K} L(\theta) = \frac{1}{(\sqrt{2\pi})^n (\hat{\sigma}^2)^{\frac{n}{2}}} e^{-\frac{n}{2}}$$

$$\text{and } \max_{\theta \in \Omega_H} L(\theta) = \frac{1}{(\sqrt{2\pi})^n (\hat{\hat{\sigma}}^2)^{\frac{n}{2}}} e^{-\frac{n}{2}}$$

. $\Rightarrow$ The LR statistic :

$$\lambda(\mathbf{x}) = \frac{max_{\theta \in \Omega_H} L(\theta)}{max_{\theta \in \Omega_H \cup \Omega_K} L(\theta)} = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} \right)^{\frac{n}{2}} = \left(\frac{SSW}{TSS} \right)^{\frac{n}{2}}.$$

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

$$\text{Here, } CR : \lambda(\mathbf{x}) < C \Leftrightarrow \frac{SSW}{TSS} < C_1$$

$$\text{or, } \frac{\sum_i \sum_j (x_{ij} - \overline{x_i})^2}{\sum_i \sum_j (x_{ij} - \overline{x_i})^2 + \sum_i n_i(\overline{x_i} - \overline{x})^2} < C_1$$

$$\Leftrightarrow \frac{\sum_i n_i(\overline{x_i} - \overline{x})^2}{\sum_i \sum_j (x_{ij} - \overline{x_i})^2} > C_2 \ (\textit{on simplification})$$

$$\text{or, } \frac{\sum_i n_i(\overline{x_i} - \overline{x})^2/(k-1)}{\sum_i \sum_j (x_{ij} - \overline{x_i})^2/(n-k)} > C_3$$

$$\text{i.e., } \frac{SSB/(k-1)}{SSW/(n-k)} > C_3.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Actually the problem that we have considered here is the ANALYSIS OF VARIANCE problem is its simplest form with one-way classified data under fixed effect model.

$\Rightarrow$ the ultimate statistic what we get here is :

$$F_H = \frac{SSB/(k-1)}{SSW/(n-k)} = \frac{MSB}{MSW} \overset{H}{\sim} F_{k-1,n-k}.$$

We reject H if and only if $F_H > C_3$ i.e., if and only if the observed value of F_H is significantly large.
From size α condition, we get : $C_3 = F_{\alpha;k-1,n-k}$.

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

Application 2.

- Next, suppose we want to test $H : \sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ (say) versus K : *not H*. Now $\theta = (\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$.
- $\Rightarrow \ \Omega_H = \{\theta : -\infty < \mu_i < \infty, i = 1, \ldots, k; \ \sigma_i^2 = \sigma^2 \ \forall \ i = 1, \ldots, k; \ \sigma > 0\}$
- and $\Omega_H \cup \Omega_K = \{\theta : -\infty < \mu_i < \infty; \ \sigma_i^2 > 0 \ \forall \ i = 1, \ldots, k\}$.
- Under $\Omega_H \cup \Omega_K$,

$$L(\theta) = \frac{1}{(\sqrt{2\pi})^n \prod_{i=1}^k (\sigma_i^2)^{\frac{n_i}{2}}} e^{-\frac{1}{2}\sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(x_{ij}-\mu_i)^2}{\sigma_i^2}}.$$

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

Here, we can easily derive the MLE of μ_i to be $\overline{x_i}$ and MLE of σ_i^2 as $s_i^2 = \frac{1}{n_i}\sum_j (x_{ij} - \overline{x_i})^2 \rightarrow$ the sample variance with divisor n_i for ith population, $i = 1, \ldots, k$.

$$\Rightarrow \quad sup_{\theta \in \Omega_H \cup \Omega_K} L(\theta) = \frac{1}{(\sqrt{2\pi})^n \prod_{i=1}^k (\sigma_i^2)^{\frac{n_i}{2}}} e^{-\frac{n}{2}}.$$

Again, under Ω_H,

$$L(\theta) = \frac{1}{(\sqrt{2\pi})^n (\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Here, MLE of μ_i is as usual $\overline{x_i}$, but now MLE of σ^2 is $\hat{\sigma^2} = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} \rightarrow$ a weighted average of sample variances, weights being proportional to the sample sizes.

$\Rightarrow$ Now, $sup_{\theta \in \Omega_H} L(\theta) = \frac{1}{(\sqrt{2\pi})^n \hat{\sigma^2}^{\frac{n}{2}}} e^{-\frac{n}{2}}$.

$\Rightarrow$ the LR statistic :

$$\lambda(\mathbf{x}) = \prod_{i=1}^{k} \left(\frac{s_i^2}{\hat{\sigma^2}} \right)^{\frac{n_i}{2}} = \frac{\prod_{i=1}^{k} (s_i^2)^{\frac{n_i}{2}}}{\left(\frac{1}{n} \sum_{i=1}^{k} n_i s_i^2\right)^{\frac{n}{2}}}.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

The Critical Region : $\lambda(\mathbf{x}) < C$

$$\Leftrightarrow \frac{\left(\prod_{i=1}^{k}(s_i^2)^{n_i}\right)^{\frac{1}{n}}}{\frac{1}{n}\sum_{i=1}^{k} n_i s_i^2} < C_1$$

or, $\dfrac{\textit{weighted GM of } s_i^2}{\textit{weighted AM of } s_i^2} < C_1$, weights being sample size.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

- Since an unbiased estimator of σ_i^2 is $s_i'^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2$, the sample variance with divisor $(n_i - 1)$ for ith population, replacing s_i^2 by $s_i'^2$, the CR can alternatively be written as : $\frac{\textit{weighted GM of } s_i'^2}{\textit{weighted AM of } s_i'^2} < C_2$, weights being $\nu_i = n_i - 1$ and $\sum_i \nu_i = \nu$(say) $=$ total weight.
- an improved LR statistic (in terms of $s_i'^2$) is :

$$\lambda^* = \frac{[(s_1'^2)^{\nu_1} \ldots (s_k'^2)^{\nu_k}]^{\frac{1}{\nu}}}{\frac{\nu_1 s_1'^2 + \cdots + \nu_k s_k'^2}{\nu}}.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

For computational purposes, it is better to use the statistic : $M = -\nu log_e(\lambda^*)$, which was suggested by Bartlett.

$$\Rightarrow \quad M = \nu log_e \left\{ \frac{\nu_1 {s_1'}^2 + \cdots + \nu_k {s_k'}^2}{\nu} \right\} - \nu_1 log_e {s_1'}^2 - \cdots - \nu_k log_e {s_k'}^2.$$

Now, let us consider the following two approximations to the asymptotic null distributions of test statistic.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

- **1st approximation :** For large $\nu_1, \ldots, \nu_k$, the asymptotic null distribution of M is central χ^2 with degrees of freedom $2k - (k+1) = k - 1$.
- **2nd approximation :** If ν_i's are not large, i.e., in small sample case, Bartlett proved that the statistic

$$M' = \frac{M}{1 + \frac{C_1}{3(k-1)}},$$

where, $C_1 = (\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\nu})$ has the null distribution tending to central χ^2 with faster rate than M itself.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Application 3. Let $Y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$ independently for all $i = 1, \ldots, p$ and for all $j = 1, \ldots, q$. So the data are framed as two-way classified data as if with a factor A having p levels with additional effects α_is and with another factor B having q levels with additional effects β_js.

Also assume $\sum_{i=1}^{p} \alpha_i = 0 \,, \forall\, j$ and $\sum_{j=1}^{q} \beta_j = 0 \,, \forall\, i$.

$\implies$ the parameter $\theta = (\mu, \alpha_i, \beta_j, \sigma^2 \,;\, i = 1, \ldots, p \,,\, j = 1, \ldots, q)$.

Suppose we want to test $H_1 : \alpha_i = 0$ for all i versus K_1 : not H_1 , (for all j). And $H_2 : \beta_j = 0$ for all j versus K_2 : not H_2 , (for all i.)
So $H_1 \iff$ factor A not significant and $H_2 \iff$ factor B not significant.

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

Under $\Omega_{H_1} \cup \Omega_{K_1}$, the loglikelihood of θ :

$$\ell(\theta) = \text{Const.} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \mu - \alpha_i - \beta_j)^2\,;\; n = pq$$

So here the ML estimates are equivalent to Least Square estimates. Hence here the ML estimate of (μ, α_i, β_j) is $\left(\overline{y}_{..}, \overline{y}_{i.} - \overline{y}_{..}, \overline{y}_{.j} - \overline{y}_{..}\right)$, where

$$\overline{y}_{..} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}\,(\text{grand mean})\,,\; \overline{y}_{i.} = \frac{1}{q}\sum_{j=1}^{q} y_{ij}\,,\; \overline{y}_{.j} = \frac{1}{p}\sum_{i=1}^{p} y_{ij}.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Thus the ML estimate of σ^2 becomes

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..} - (\bar{y}_{i.} - \bar{y}_{..}) - (\bar{y}_{.j} - \bar{y}_{..}))^2$$

which, on simplification, becomes

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\left\{\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..})^2 - q\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2 - p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2\right\} \\ &= \frac{1}{n}\{TSS - SSA - SSB\} = \frac{1}{n}SSE \end{aligned}$$

where TSS = Total Sum of Squares , SSA = Sum of Squares Due to Factor A, SSB = Sum of Squares Due to Factor B and SSE = Sum of Squares Due to Error

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Under Ω_{H_1}, the loglikelihood of θ :

$$\ell(\theta) = \text{Const.} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \mu - \beta_j)^2.$$

For the same reason here also the ML estimates of μ and β_j remain unchanged and as a result now the ML of σ^2 becomes

$$\begin{aligned}
\hat{\hat{\sigma}}^2 &= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..} - (\bar{y}_{.j} - \bar{y}_{..}))^2 \\
&= \frac{1}{n}\left\{\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..})^2 - p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2\right\} \text{ (on simplification)} \\
&= \frac{1}{n}\{TSS - SSB\} = \frac{1}{n}\{SSE + SSA\}
\end{aligned}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

$\implies \sup\limits_{\Omega_{H_1} \cup \Omega_{K_1}} L(\theta) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\{-n/2\}$, and similarly

$$\sup_{\Omega_{H_1}} L(\theta) = \frac{1}{\left(2\pi\hat{\hat{\sigma}}^2\right)^{n/2}} \exp\{-n/2\}$$

$\implies$ the LR criterion is $\lambda(\mathbf{y}) = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2}\right)^{n/2}$.

The critical region is : $\lambda(\mathbf{y}) < \lambda \iff \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} < \lambda_1$.

Or $\frac{SSE+SSA}{SSE} > \lambda_2$ i.e. equivalent to $\frac{SSA}{SSE} > \lambda_3$.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Now, from orthogonal splitting (due to Cochran's Theorem) we know that $\frac{SSE}{\sigma^2}$ and $\frac{SSA}{\sigma^2}$ are independenlty distributed with $\frac{SSE}{\sigma^2} \sim \chi^2_{(p-1)(q-1)}$ and $\frac{SSA}{\sigma^2} \overset{H_1}{\sim} \chi^2_{(p-1)}$.

Hence the CR can equivalently be written as $F_{H_1} = \frac{\frac{SSA}{\sigma^2}/p-1}{\frac{SSE}{\sigma^2}/(p-1)(q-1)} > c$

where $F_{H_1} \overset{H_1}{\sim} F_{p-1,(p-1)(q-1)}$ and c is such that $P_{H_1}[CR] = \gamma$ (given significance level)

$\Rightarrow \; c = F_{\gamma;p-1,(p-1)(q-1)}$: upper 100γ % point of $F_{p-1,(p-1)(q-1)}$ distribution.

Similarly we can perform LRT for H_2 versus K_2.

LRT : Applications under k ($>$ 2) comparable single distributions and Multivariate Distribution

Application 4. Next, suppose $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be i.i.d. $N_p(\boldsymbol{\mu}, \Sigma)$. Here $\theta = (\boldsymbol{\mu}, \Sigma)$.

Let us test $H : \boldsymbol{\mu} = \boldsymbol{\mu_0}$ (known) versus $K : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$.

$\Rightarrow \ \Omega_H = \{(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} = \boldsymbol{\mu_0}, \Sigma \textit{ is } p.d.\}$ and
$\Omega_H \cup \Omega_K = \{(\boldsymbol{\mu}, \Sigma) : -\infty < \mu_i < \infty, \Sigma \textit{ is } p.d.\}$.

Under $\Omega_H \cup \Omega_K$, the likelihood is :

$$L(\theta) = \frac{1}{(2\pi)^{\frac{np}{2}}(|\Sigma|)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}_\alpha - \boldsymbol{\mu})} .$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

$$\Rightarrow \quad L(\theta) = \frac{1}{(\sqrt{2\pi})^{np}|\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2}\{n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) + \sum_{\alpha=1}^{n}(\mathbf{x}_\alpha-\bar{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_\alpha-\bar{\mathbf{x}})\}}$$

where, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{\alpha=1}^{n}\mathbf{x}_\alpha \;\rightarrow\;$ sample mean vector.
or,

$$L(\theta) = \frac{1}{(\sqrt{2\pi})^{np}|\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2}\{n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) + n\, tr.\Sigma^{-1}V\}}$$

where, $V = \frac{1}{n}\sum_{\alpha}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})' \;\rightarrow\;$ sample variance - covariance matrix with divisor n.

$\Rightarrow$ MLE of $\boldsymbol{\mu}$ is $\bar{\mathbf{x}}$ and of Σ is V.

$$\Rightarrow \quad \max_{\theta\in\Omega_H\cup\Omega_K} L(\theta) = \frac{1}{(\sqrt{2\pi})^{np}|V|^{\frac{n}{2}}} e^{-\frac{np}{2}} \; .$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Similarly under Ω_H,

$$L(\theta) = \frac{1}{(\sqrt{2\pi})^{np}|\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha-\boldsymbol{\mu}_\mathbf{0})'\Sigma^{-1}(\mathbf{x}_\alpha-\boldsymbol{\mu}_\mathbf{0})}$$
$$= \frac{1}{(\sqrt{2\pi})^{np}|\Sigma|^{\frac{n}{2}}} e^{-\frac{n}{2}tr.\Sigma^{-1}V_0}$$

where, $V_0 = \frac{1}{n}(\mathbf{x}_\alpha - \boldsymbol{\mu}_\mathbf{0})'\Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}_\mathbf{0})$.

$$\Rightarrow \; l(\theta) = log_e L(\theta) = const. + \frac{n}{2} log_e|\Sigma^{-1}V_0| - \frac{n}{2} tr.\Sigma^{-1}V_0$$
$$= const. + \frac{n}{2}\sum_{i=1}^{p} log_e(\nu_i) - \frac{n}{2}\sum_{i=1}^{p}\nu_i = \psi(\boldsymbol{\nu}) \; (say),$$

where, ν_i's are the eigen values of $\Sigma^{-1}V_0$.

LRT : Applications under k (>2) comparable single distributions and Multivariate Distribution

Now, $\frac{\partial \psi(\boldsymbol{\nu})}{\partial \nu_j} = 0 \Rightarrow \frac{n}{2\nu_j} - \frac{n}{2} = 0$ i.e., $\nu_j = 0, \; \forall j = 1, \ldots, p$.
and $\left.\frac{\partial \psi(\boldsymbol{\nu})}{\partial \nu_j}\right|_{\nu_j = 1} = -\frac{n}{2} < 0$, $\frac{\partial^2 \psi(\boldsymbol{\nu})}{\partial \nu_j \partial \nu_k} = 0 \; \forall j \neq k$.

$\Rightarrow \psi(\boldsymbol{\nu})$ and hence, $l(\theta)$ gets maximised for $\nu_i = 1, \; \forall \, i$, i.e., for $\Sigma^{-1} V_0 = I_p$.

$\Rightarrow$ under Ω_H, the MLE of Σ is $\hat{\Sigma} = V_0$.

$$\Rightarrow \max_{\theta \in \Omega_H} L(\theta) = \frac{1}{(\sqrt{2\pi})^{np} |\hat{\Sigma}|^{\frac{n}{2}}} e^{-\frac{np}{2}} \; . \tag{1}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

$\Rightarrow$ the LR statistic is : $\Lambda = \dfrac{\max_{\theta \in \Omega_H} L(\theta)}{\max_{\theta \in \Omega_H \cup \Omega_K} L(\theta)} = \left(\dfrac{|V|}{|\hat{\Sigma}|} \right)^{\frac{n}{2}}$.

The equivalent statistic is : $\Lambda^{\frac{2}{n}} = \dfrac{|V|}{|\hat{\Sigma}|}$ is called Wilk's Lambda.

CR : $\Lambda < C \;\Leftrightarrow\; \Lambda^{\frac{2}{n}} < C_1$
or, $\dfrac{|V|}{|\hat{\Sigma}|} < C_1$, where, $\hat{\Sigma} = \frac{1}{n} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \boldsymbol{\mu}_0)'$.
Now,

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\mathbf{x}_\alpha - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \\
&= \frac{1}{n} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})' + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'
\end{aligned}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Hence, $\hat{\Sigma} = V + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'$.
Now, under Ω_H,

$$\mathbf{y} = \sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim N_p(\mathbf{0}, \Sigma).$$

$$\text{and } S = \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})' \sim Wishart_p(\Sigma, n-1) \text{ independently.}$$

$\Rightarrow$ under Ω_H, $T^2 = (n-1)\mathbf{y}' S^{-1}\mathbf{y} \sim T^2_{p,n-1}$ (Central T^2 distribution with dimension p and degrees of freedom $n-1$).
Now,

$$V = \frac{S}{n} \;\Rightarrow\; T^2 = n(n-1)\mathbf{y}' V^{-1}\mathbf{y}.$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

Hence, $\hat{\Sigma} = V + \frac{\mathbf{y}\mathbf{y}'}{n}$. So, the CR is : $\frac{|V|}{|\hat{\Sigma}|} < C_1 \Rightarrow \frac{|V|}{|V + \frac{\mathbf{y}\mathbf{y}'}{n}|} < C_1$.

As, $|V + \frac{\mathbf{y}\mathbf{y}'}{n}| = (1 + \frac{\mathbf{y}'V^{-1}\mathbf{y}}{n})$, so the CR is : $\frac{1}{1 + \frac{T^2}{n^2(n-1)}} < C_1$.

$\Leftrightarrow T^2 > C_2$, where from size α condition :

$$C_2 = T^2_{p,n-1}(\alpha) = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

.

where, $T^2_{p,n-1}(\alpha) \rightarrow$ upper $100\alpha\%$ point of T^2 distribution with dimension p and degrees of freedom $n-1$.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

TRY YOURSELF!

13\. 1. Let $Y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$ independently for all $i = 1, \ldots, p$ and for all $j = 1, \ldots, q$. So the data are framed as two-way classified data as if with a factor A having p levels with additional effects α_is and with another factor B having q levels with additional effects β_js.

Also assume $\sum_{i=1}^{p} \alpha_i = 0 \,, \forall\, j$ and $\sum_{j=1}^{q} \beta_j = 0 \,, \forall\, i$.

If σ is known, show that the LRT for ANOVA is performed by the χ^2 test statistic.

LRT : Applications under k (>2) comparable single distributions and Multivariate Distribution

13. 2. Let $\{X_{11}, X_{21}, \ldots, X_{r1}\}, \ldots, \{X_{1c}, X_{2c}, \ldots, X_{rc}\}$ be independent multinomial random variables with parameters $(n_1; p_{11}, p_{21}, \ldots, p_{r1}), \ldots, (n_c; p_{1c}, p_{2c}, \ldots, p_{rc})$ respectively. Let $X_{i.} = \sum_{j=1}^{c} X_{ij}$ and $\sum_{j=1}^{c} n_j = n$. Show that the LRT for testing $H : p_{ij} = p_i \ , \forall\, i = 1, \ldots, r;$ for all $j = 1, 2, \ldots, c$ against all alternatives K can be based on the statistic

$$\lambda(\mathbf{x}) = \prod_{i=1}^{r} \left(\frac{X_{i.}}{n} \right)^{X_{i.}} \Big/ \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{X_{ij}}{n_j} \right)^{X_{ij}}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

TUTORIAL DISCUSSION :

Overview to the problems from MODULE 13 ...

M13.1. Since the basic set up of the problem and hypotheses to be tested are same as in **Application 3.** so we will use same notations and formule in this problem also.
The ML estimates remain same as in **Application 3.** except of σ which is known ($= \sigma_0$, say) here. So

$$\begin{aligned}
\sup_{\Omega_{H_1}} L(\theta) &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i,j}(y_{ij} - \bar{y}_{..} - (\bar{y}_{.j} - \bar{y}_{..}))^2\right\} \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}(TSS - SSB)\right\}
\end{aligned}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

$$\text{Similarly,} \quad \sup_{\Omega_{H_1} \cup \Omega_{K_1}} L(\theta) = \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{ -\frac{\sum_{i,j} (y_{ij} - \bar{y}_{..} - (\bar{y}_{i.} - \bar{y}_{..}) - (\bar{y}_{.j} - \bar{y}_{..}))^2}{2\sigma_0^2} \right\}$$

$$= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{ -(TSS - SSA - SSB)/2\sigma_0^2 \right\}$$

If $\lambda(\mathbf{y})$ denotes LR criterion then here

$$\ln \lambda(\mathbf{y}) = -\frac{1}{2\sigma_0^2}[TSS - SSB] + \frac{1}{2\sigma_0^2}[TSS - SSA - SSB] = -\frac{SSA}{2\sigma_0^2}$$

$$\implies -2 \ln \lambda(\mathbf{y}) = \frac{SSA}{\sigma_0^2} \overset{H_1}{\sim} \chi^2_{p-1}$$

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

[**Justification :** From sampling distribution we can write
$Z_i = \overline{Y}_{i.} \overset{H_1}{\sim} N(\mu, \frac{\sigma_0^2}{q} = \sigma^*)$ independently $\forall\, i = 1, \ldots, p$.
Hence, again from the well-known result of the sampling distribution on normal population we get

$$\frac{\sum_{i=1}^{p}(Z_i - \overline{Z})^2}{\sigma^*} \overset{H_1}{\sim} \chi^2_{p-1} \quad i.e. \quad \frac{q\sum_{i=1}^{p}(Y_{i.} - \overline{Y}_{..})^2}{\sigma_0^2} \overset{H_1}{\sim} \chi^2_{p-1}$$

$$Or \ \frac{SSA}{\sigma_0^2} \overset{H_1}{\sim} \chi^2_{p-1} \ ]$$

So the CR : $\lambda(\mathbf{y}) < \lambda \Leftrightarrow -2\ln\lambda(\mathbf{y}) > c$ i.e. $\chi^2_{H_1} = \frac{SSA}{\sigma_0^2} > c$, where c is determined as $\chi^2_{\gamma;p-1}$ from the size-γ condition. Similarly, we perform test of H_2 versus K_2.

LRT : Applications under k (> 2) comparable single distributions and Multivariate Distribution

M13. 2. Without loss of generality here we assume singular multinomial distribution. So we can use

$$p_{1j} + p_{2j} + \ldots + p_{rj} = 1 \; \forall \, j \text{ and under } H, \; p_1 + p_2 + \ldots + p_r = 1$$

From discussion on MLE in Module 5, we get that under $\Omega_H \cup \Omega_K$ the ML estimate of p_{ij} is $\frac{x_{ij}}{n_j}$. Thus

$$\sup_{\Omega_H \cup \Omega_K} L(\theta) = \left(\prod_{j=1}^{c} \frac{n_j!}{x_{1j}! \ldots x_{rj}!} \right) \prod_{j=1}^{c} \prod_{i=1}^{r} \left(\frac{x_{ij}}{n_j} \right)^{x_{ij}}$$

On the other hand the likelihood under Ω_H is

$$L(\theta) = \left(\prod_{j=1}^{c} \frac{n_j!}{x_{1j}! \ldots x_{rj}!} \right) p_1^{\sum_j x_{1j}} \ldots p_r^{\sum_j x_{rj}} \; ; \; 0 < p_i < 1.$$

LRT : Applications under k (>2) comparable single distributions and Multivariate Distribution

Then following the conventional way of finding MLE under multinomial probability model (*discussed in Module 5*) we get MLE of p_i as

$$\frac{\sum_j x_{ij}}{n} = \frac{x_{i.}}{n} \text{ (Readers! Try yourselves)}$$

$$\implies \sup_{\Omega_H} L(\theta) = \left(\prod_{j=1}^{c} \frac{n_j!}{x_{1j}!\ldots x_{rj}!}\right) \prod_{i=1}^{r} \left(\frac{x_{i.}}{n}\right)^{x_{i.}}.$$

Thus finally the LR criterion will be

$$\lambda(\mathbf{x}) = \frac{\sup_{\Omega_H} L(\theta)}{\sup_{\Omega_H \cup \Omega_K} L(\theta)} = \prod_{i=1}^{r} \left(\frac{x_{i.}}{n}\right)^{x_{i.}} / \prod_{j=1}^{c}\prod_{i=1}^{r} \left(\frac{x_{ij}}{n_j}\right)^{x_{ij}}.$$

Some Asymptotically Equivalent Tests

Module 14

Saurav De

Department of Statistics
Presidency University

Now let us consider the following tests having the similar asymptotic behaviour.

Start with the following asymptotic aspects.

We know from definition of likelihood ratio criterion that

$$\lambda(x) = \frac{L_x(\theta_0)}{L_x(\hat{\theta}_n)}$$

where x denotes data from random sampling, $\hat{\theta}_n$: the MLE of θ, θ_0: hypothesised value of θ.

$\implies -2\ln\lambda(x) = 2\left[l_x(\hat{\theta}_n) - l_x(\theta_0)\right]$, where l_x denotes the loglikelihood.

From Taylor's expansion we can write

$$l_x(\theta_0) \approx l_x(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n) l'_x(\hat{\theta}_n) + \frac{(\hat{\theta}_n - \theta_0)^2}{2} l''_x(\hat{\theta}_n)$$

Now $l'_x(\hat{\theta}_n) = 0$ as $\hat{\theta}_n$: MLE of θ.

$\implies \; l_x(\theta_0) - l_x(\hat{\theta}_n) \approx \frac{(\hat{\theta}_n - \theta_0)^2}{2} l''_x(\hat{\theta}_n)$

Or $-2\ln\lambda(x) = 2\left[l_x(\hat{\theta}_n) - l_x(\theta_0)\right] \overset{a}{\approx} \left\{\sqrt{n}(\hat{\theta}_n - \theta_0)\right\}^2 \left(-\frac{1}{n} l''_x(\hat{\theta}_n)\right)$

Now use the fact that

$$-\frac{1}{n} l_x''(\hat{\theta}_n) = \frac{1}{n}\sum - \frac{\partial^2}{\partial\theta^2} f_\theta(x_i) \mid_{\theta=\hat{\theta}_n} \approx i(\hat{\theta}_n).$$

Hence we get

$$Q_{LR} = -2\ln\lambda(x) = 2\left[l_x(\hat{\theta}_n) - l_x(\theta_0)\right] \overset{a}{\approx} \left\{\sqrt{n}(\hat{\theta}_n - \theta_0)\right\}^2 i(\hat{\theta}_n)$$

Again expanding with respect to θ_0, we get from another Tailor's expansion

$$l'_x(\hat{\theta}_n) \approx l'_x(\theta_0) + (\hat{\theta}_n - \theta_0) l''_x(\theta_0)$$

But $l'_x(\hat{\theta}_n) = 0$ as $\hat{\theta}_n$: MLE of θ.

$\implies \hat{\theta}_n - \theta_0 \approx -\frac{l'_x(\theta_0)}{l''_x(\theta_0)}$

Or

$$\left\{\sqrt{n}(\hat{\theta}_n - \theta_0)\right\}^2 \approx n\frac{(l'_x(\theta_0))^2}{(l''_x(\theta_0))^2} = \frac{(l'_x(\theta_0))^2}{n(-\frac{1}{n}l''_x(\theta_0))^2} \stackrel{a}{\approx} \frac{\{l'_x(\theta_0)\}^2}{ni^2(\theta_0)}$$

Or we can write

$$\left\{\sqrt{n}(\hat{\theta}_n - \theta_0)\right\}^2 i(\hat{\theta}_n) \stackrel{a}{\approx} \frac{\{l'_x(\theta_0)\}^2}{ni(\theta_0)} \quad \text{under } \theta_0 \ldots\ldots (*)$$

Alternative way: follow the approach of iterative MLE and take initial approximation as θ_0 itself.

- Hence we can alternatively say that
  $Q_{LR} = -2\ln\lambda(X) = 2\left[l_X(\hat{\theta}_n) - l_X(\theta_0)\right] \overset{a}{\approx} \frac{\{l_x'(\theta_0)\}^2}{ni(\theta_0)}$ under θ_0.

- Also $\frac{1}{\sqrt{n}} l_x'(\theta_0) = \frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial\theta} f_\theta(x_i) \mid_{\theta_0} \xrightarrow{\mathcal{D}} N(0, i(\theta_0)) \ldots\ldots (**)$

  [Justification : Define $Y_i = \frac{\partial}{\partial\theta} f_\theta(x_i) \mid_{\theta_0}$. So Y_i for $i = 1, 2, \ldots$ are iid (as X_is are iid) with $E(Y_1) = 0$; $V(Y_1) = i(\theta_0)$

  Hence by CLT $\frac{1}{\sqrt{n}} \sum Y_i = \sqrt{n}\overline{Y} \xrightarrow{\mathcal{D}} N(0, i(\theta_0))$]

Based on the above asymptotic equivalence and asymptotic normality, we have the following two popular tests.

WALD'S TEST

Single parameter(θ) case: (for θ : real valued) Suppose the Fisher's information $i(\theta)$ exists at all θ. Then one should reject $H : \theta = \theta_0$ if the observed value of the statistic

$$Q_W = n(\hat{\theta}_n - \theta_0)^2 i(\hat{\theta}_n)$$

is significantly large. This Q_W is known as the Wald test statistic for single parameter testing $H : \theta = \theta_0$.

Provided • certain regularity conditions for the asymptotic normality of the MLE $\hat{\theta}_n$ hold and • assuming that $i(\theta)$ is continuous in the neighbourhood of θ_0, $(*)$ and $(**)$ directly imply, under $H : \theta = \theta_0$,

$$Q_W \xrightarrow{\mathcal{D}} \chi^2_1.$$

Hence for large n, $H : \theta = \theta_0$ will be rejected at level α if

$$Q_W = n(\hat{\theta}_n - \theta_0)^2 i(\hat{\theta}_n) > \chi^2_{1;\alpha}$$

Multi-parameter($\boldsymbol{\theta}$) case: (for $\boldsymbol{\theta}$: s-component vector valued) Suppose $\hat{\boldsymbol{\theta}}_n$: MLE of $\boldsymbol{\theta}$ and the Fisher information matrix $I(\boldsymbol{\theta})$ exists for all $\boldsymbol{\theta}$. Then the null hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ should not be accepted if the value of the statistic

$$Q_W = \left\{\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\}' I(\hat{\boldsymbol{\theta}}_n) \left\{\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\}$$

is significantly high.

Assuming $I(\boldsymbol{\theta})$ is smooth at $\boldsymbol{\theta}_0$ and the regularity conditions for the asymptotic multivariate normality for $\hat{\boldsymbol{\theta}}_n$ hold, under $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$Q_W = \left\{\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\}' I(\hat{\boldsymbol{\theta}}_n) \left\{\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\} \xrightarrow{\mathcal{D}} \chi^2_s.$$

$\Rightarrow$ for large n, reject H at level α if $Q_W > \chi^2_{s;\alpha}$.

RAO'S SCORE TEST

Single parameter(θ) case: Suppose $l_x(\theta)$ is differentiable with respect to θ for each x. Then $\frac{\partial}{\partial\theta} l_x(\theta) = l'_x(\theta)$ is called score function of θ.
Also suppose the Fisher information $i(\theta)$ exists and is positive at θ_0.

Now if θ_0 is the true value of θ, it will be close enough to $\hat{\theta}_n$. As a result $l'_x(\theta_0)$ should be close to 0 (since $l'_x(\hat{\theta}_n) = 0$).
Thus signifivantly high value of $\frac{\{l'_x(\theta_0)\}^2}{n\ i(\theta_0)}$ leaves evidence against the null hypothesis $H : \theta = \theta_0$.

The statistic $Q_S = \frac{\{l'_x(\theta_0)\}^2}{n\ i(\theta_0)}$ is called Rao score statistic under real valued parametric case.

Also , under $H : \theta = \theta_0$ the asymptotic distribution of $\frac{\{l_x'(\theta_0)\}^2}{n\; i(\theta_0)}$ is χ^2 with 1 degree of freedom.

[**Justification:** $\frac{\{l_x'(\theta_0)\}^2}{n\; i(\theta_0)} \overset{a}{\approx} \left\{\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right\}^2 i(\hat{\theta}_n)$.

And $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N(0, \frac{1}{i(\theta_0)})$

i.e. $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\sqrt{i(\theta_0)} \xrightarrow{\mathcal{D}} N(0, 1)$ (*using asymptotic distribution of univariate MLE*). Hence justified.]

Thus for large n, under real-valued parameter θ, Rao's score test:

- reject H at level α if $\frac{\{l_x'(\theta_0)\}^2}{n\ i(\theta_0)} > \chi^2_{1;\alpha}$

Multi-parameter($\boldsymbol{\theta}$) case: Let $\boldsymbol{\theta}$ be $s(>1)$ component. Suppose

- l_x is differentiable partially under the integral sign with respect to each component θ_j of $\boldsymbol{\theta}$ for every x,
- the Fisher information matrix $I(\boldsymbol{\theta})$ exists and is invertible at $\boldsymbol{\theta}_0$, and
- the support of the parent distribution is independent of $\boldsymbol{\theta}$.

Then under $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, (*using the asymptotic multivariate normality of vector-valued* $\hat{\boldsymbol{\theta}}_n$)

$$Q_S = \mathbf{S}' I^{-1}(\boldsymbol{\theta}_0)\mathbf{S} \xrightarrow{\mathcal{D}} \chi^2_s$$

where $S_j = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta_j} l_x(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

The vector $\mathbf{S}$ is called the score vector.

Following the same idea as in *Single parameter(θ) case* for large n, reject H at level α if

$$Q_S = \mathbf{S}' I^{-1}(\boldsymbol{\theta}_0)\mathbf{S} > \chi^2_{s;\alpha}$$

Rao's score test

- is also known as the **Lagrange multiplier test** in econometrics,
- is frequently meant for testing a simple nul hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and
- provides the most powerful test when the true value of $\boldsymbol{\theta}$ lies in the close vicinity of of $\boldsymbol{\theta}_0$.

Why the Rao's score test most powerful in locality of θ_0:

Neyman-Pearson lemma always provides the most powerful test and the the test can be expressed as

$\frac{L(\theta_0+h|x)}{L(\theta_0|x)} \geq K$

$\Leftrightarrow \log L(\theta_0 + h \mid x) - \log L(\theta_0 \mid x) \geq \log K \qquad (*)$

Some Asymptotically Equivalent Tests

Also from Taylor series expansion we can write

$$\log L(\theta_0 + h \mid x) \approx \log L(\theta_0 \mid x) + h. \left(\frac{\partial}{\partial \theta} \log \log L(\theta \mid x) \right)_{\theta = \theta_0} \qquad (**)$$

[*ignoring the higher order terms as these are negligible if h is very small i.e. if $\theta = \theta_0 + h$ is close to θ*]

Hence from $(*)$ and $(**)$ we can say that

$$h.\left(\frac{\partial}{\partial\theta}\log\log L(\theta \mid x)\right)_{\theta=\theta_0} \geq \log K$$

or in other way $\frac{\{l_x'(\theta_0)\}^2}{n\ i(\theta_0)} > c$, where c is chosen from the size(α) condition. In particular, for large n, $c = \chi^2_{1;\alpha}$.

Note that, under simple null hypotheses

- Q_S uses less assumptions compared to Q_W and Q_{LR} for χ^2 approximation
- Q_S statistic does not require computation of the maximum likelihood estimate of θ.

  So these features are advantageous because
- the test is applicable even when the unconstrained maximum likelihood estimate is a boundary point in the parameter space, and
- the null distribution of Q_S can be approximated by χ^2 in wide applications.

Note.

- Like Q_{LR}, Q_W and Q_S can also be framed for testing composite nulls of the form $H : \gamma(\boldsymbol{\theta}) = (\gamma_1(\boldsymbol{\theta}), \ldots, \gamma_q(\boldsymbol{\theta})) = 0$. Q_S statistic, now, depend on the restricted MLE of $\boldsymbol{\theta}$ under H. However in this case also, the asymptotic χ^2_q distribution under H is maintained by each of these three statistics.
- In literature of statistics, LR test, Wald test and Rao's score test, can be referred to as the Holy Trinity due to their wide applicability.
- All these tests are asymptotically equivalent at least to the first order of asymptotics, but
- may differ to some extent in the second (or higher) order properties.

  As a result of fact,
- no one is uniformly superior to the others.

Likelihood Ratio, Wald and Score Confidence Sets

Using the usual duality (i.e. the equivalence) between CRs of testing of hypothesis and confidence sets, the acceptance region (complementary set of CR) of a test with size α can be inverted to a confidence set with confidence coefficient at least $(1-\alpha)$. This method is known as the inversion of a test.

Let $A(\theta_0)$ denote the acceptance region of a size-α test for testing $H : \theta = \theta_0$. Inverting the set $A(\theta_0)$ we get the set

$$S(x) = \{\theta_0 : x \in A(\theta_0)\} \ni \ P_{\theta_0}[S(X) \ni \theta_0] \geq 1 - \alpha$$

i.e. in general $P_\theta[S(X) \ni \theta] \geq 1 - \alpha$.

$\implies S(X)$: a $100(1-\alpha)\%$ confidence set for θ.

The confidence sets obtained from the LRT, the Wald test and the Score test are respectively known as the Likelihood Ratio, Wald and Score Confidence Sets.

Note that

- Since the Wald and the Score test statistics are defined as quadratic forms, the corresponding confidence sets will be ellipsoid.
- The LR confidence set does not have any such pre-specified form and is typically more complicated.

We can illustrate the construction of these confidence sets through the following example :

Example. Suppose $x_1, \ldots, x_n$ be n realisations on Bernoulli(θ) distribution. Let $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$.

Some Asymptotically Equivalent Tests

By LRT the LR criterion will be

$$\lambda(x) = \frac{\theta_0^x(1-\theta_0)^{n-x}}{\sup\limits_{\theta}\{\theta^x(1-\theta)^{n-x}\}}\ ,\ x = \sum_{i=1}^{n} x_i$$

$$= \frac{\theta_0^x(1-\theta_0)^{n-x}}{\hat{\theta}^x\left(1-\hat{\theta}\right)^{n-x}}\ ,\ \hat{\theta} = \frac{x}{n}$$

Then the LR critical region will be $\{x : \lambda(\mathbf{x}) < \lambda\}$ where λ is $\ni$ the size of the test is α. Thus inverting the acceptance region we get the $100(1-\alpha)\%$ exact confidence set

$$S_{LR}(x) = \left\{\theta : \frac{\theta^x(1-\theta)^{n-x}}{\hat{\theta}^x\left(1-\hat{\theta}\right)^{n-x}} \geq \lambda\right\} = \{\theta : \theta^x(1-\theta)^{n-x} \geq \lambda^*\}$$

$$= \{\theta : x\log\theta + (n-x)\log(1-\theta) \geq k\}$$

As here θ is real-valued, so $S_{LR}(x)$ will provide an interval of the form $[0, u_1]$ or $[\ell_1, 1]$ or $[\ell_2, u_2]$, $0 < \ell_1, \ell_2$, $u_1, u_2 < 1$.

Unfortunately no explicit or closed form is possible for ℓ's and u's. But using R program, we can illustrate how $100(1 - \alpha)\%$ exact confidence interval can be generated by simulation technique.

Let $n = 10$, $\theta_0 = 0.6$, $\alpha = 0.05$ and the realization be $(1, 1, 0, 1, 1, 1, 1, 0, 1, 1)$.

Illustration using R : (to find the confidence interval)

R Code and Output :

```
> library(varhandle)
> n=10
> p=0.6
> lambda=NULL
> for(j in 1:1000)
{
x=rbinom(10000,n,p)
W=(x*log(p))+((n-x)*log(1-p))
W1=sort(W)
mat=as.data.frame(table(W1))
CF=NULL
sum=0
for(i in 1:length(mat[,1]))
{
sum=sum+mat[,2][i]
CF[i]=sum
}
```

R Code and Output (continued) :

```
mat1=cbind(mat,CF)
lambda[j]=unfactor(mat1[,1][which(mat1[,3]>=9500)])[1]
x=NULL
y=NULL
W=NULL
}
> View(lambda)
> # usable lambda/cut-off value
> k=mean(lambda)
> # now, to find limits of 'p'
> # realization
> x_star=1+1+0+1+1+1+1+0+1+1
> p1=seq(0.1,0.9,0.1)
> W2=(x_star*log(p1))+((n-x_star)*log(1-p1))
> check=ifelse(W2>=k,1,0)
> mat2=data.frame(p1,W2,check)
> suit=mat2$p1[which(mat2$check==1)]
```

R Code and Output (continued) :

```
> # limits of 'p'
> lower=suit[1]
> upper=suit[length(suit)]
> lower;upper
[1] 0.7
[1] 0.9
```

Here we will have $i(\theta) = \frac{1}{\theta(1-\theta)}$.

Hence the Wald test statistic will be : $Q_W = \frac{n\left(\hat{\theta}-\theta_0\right)^2}{\hat{\theta}(1-\hat{\theta})}$.

The size-α critical region from the large sample Wald test will be $\left\{\frac{n\left(\hat{\theta}-\theta_0\right)^2}{\hat{\theta}(1-\hat{\theta})} \geq \chi^2_{1;\alpha}\right\}$.

So the $100(1-\alpha)\%$ large sample Wald confidence interval of θ will be

$$
\begin{aligned}
S_W(x) &= \left\{\theta : n\left(\hat{\theta}-\theta_0\right)^2 \leq \hat{\theta}(1-\hat{\theta})\chi^2_{1;\alpha}\right\} \\
&= \left\{\theta : |\hat{\theta}-\theta| \leq \sqrt{\frac{\chi^2_{1;\alpha}\hat{\theta}(1-\hat{\theta})}{n}}\right\} \\
&= \left[\hat{\theta} - \sqrt{\chi^2_{1;\alpha}\hat{\theta}(1-\hat{\theta})/n}\,,\ \hat{\theta} + \sqrt{\chi^2_{1;\alpha}\hat{\theta}(1-\hat{\theta})/n}\right]
\end{aligned}
$$

$\longrightarrow$ the heuristic large sample confidence interval of θ.

Again here the score function $= \frac{d}{d\theta} l_x(\theta)$

$$= \frac{d}{d\theta} x \log\theta + (n-x)\log(1-\theta) = \frac{x-n\theta}{\theta(1-\theta)}$$

Hence we have $Q_S = \frac{(x-n\theta_0)^2}{n\theta_0(1-\theta_0)}$.

Similar to the Wald approach, here also the $100(1-\alpha)\%$ large sample Score confidence interval of θ will be

$$\begin{aligned} S_S(x) &= \left\{\theta : \frac{(x-n\theta)^2}{n\theta(1-\theta)} \le \chi^2_{1;\alpha}\right\} \\ &= \{\theta : \theta^2(n^2 + n\chi^2_{1;\alpha}) - \theta(2nx + n\chi^2_{1;\alpha}) + x^2 \le 0\} \end{aligned}$$

If ℓ and u $(\ell < u)$ are two roots of the quadratic equation $\theta^2(n^2 + n\chi^2_{1;\alpha}) - \theta(2nx + n\chi^2_{1;\alpha}) + x^2 = 0$ in θ, then we have

$$S_S(x) = \{\theta : (\theta - \ell)(\theta - u) \leq 0\} = \{\theta : \ell \leq \theta \leq u\} = [\ell, u]$$

N.B. Check that here

$$\ell = \frac{2nx + nc - \sqrt{n^2c^2 + 4n^2xc - 4ncx^2}}{2(n^2 + nc)} \text{ and}$$

$$u = \frac{2nx + nc + \sqrt{n^2c^2 + 4n^2xc - 4ncx^2}}{2(n^2 + nc)},$$

where $c = \chi^2_{1;\alpha}$.

MLE under Survival Data: Type I and II Censoring

Module 15

Saurav De

Department of Statistics
Presidency University

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses

MLE under Survival Data: Type I and II Censoring

- Censoring in particular is a key issue in survival analysis.
- Censoring distinguishes survival analysis from regular statistical problems.
- Censoring is when an observation is incomplete due to some random cause.
- The cause of censoring is usually dependent on the event of interest.

MLE under Survival Data: Type I and II Censoring

Censoring differs from truncation in that the incomplete nature of the observations in truncation occurs due to a systematic selection process inherent to the study design.

Based on the directions through which incompleteness in the observations comes, cencoring is of three types

- Right Censoring
- Left Censoring
- Interval Censoring

MLE under Survival Data: Type I and II Censoring

Right censoring : The most common form of censoring
Here the lifetime of an item is followed until some time at which the event (i.e. failure or death) is yet to occur; but the event takes no further part in the study after that time.

e.g. A lung cancer patient is recruited for clinical trial to test the effect of a drug on his survival from his disease.
But he died in a car accident after T years of his disease.
$\implies$ his survival with lung cancer is at least T years, but exact years can not be known.
$\implies$ right censored.

MLE under Survival Data: Type I and II Censoring

Left censoring : This occurs when the event of interest has already taken place at the time of observation; but the exact time of occurrance of the event is not known.

e.g.

- Onset of an asymptomatic illness, like Brain Cancer
- Infection with a sexually transmitted disease like HIV / AIDS

MLE under Survival Data: Type I and II Censoring

Interval Censoring :

- Here the exact time of the occurrance of the event is not known precisely, but an interval bounding this time is known
- In case the interval is too short (e.g. 1 day or 1 hr etc.) the common practice is to ignore the interval censoring and to set one end-point of the interval consistently

e.g.

- failure of a machine during Chinese New Year celebration
- Infection with a sexually transmitted disease like HIV / AIDS in between two annual check-up

MLE under Survival Data: Type I and II Censoring

Depending on how censoring mechanism will work, there are three broad types of censoring

• Type I Censoring • Type II Censoring • Random Censoring

We will discuss in brief

- above three types in right censoring form
- the MLEs of the corresponding parameters under the survivorship probability models

MLE under Survival Data: Type I and II Censoring

Type II censoring:

- Suppose n random sample units are set on life-testing experimentation
- But due to some reasons the experiment terminates after smallest r readings
- Let these be denoted by the order statistics $T_{(1)}, \ldots, T_{(r)}$.
- Here integer r is prefixed i.e. nonrandom.

MLE under Survival Data: Type I and II Censoring

Since the remaining $n - r$ random sample values are atleast as high as $T_{(r)} \implies$ the sampling scheme is a censored one.

Such a censoring is known as Type II censoring.

Type II censoring are frequently used in life-testing experiments.

- Here say total of n items are placed on test.
- Now instead of continuing until all n items get spared, suppose the experimenter waits just for the first r failures.
- Such test saves both time and money.

MLE under Survival Data: Type I and II Censoring

Let T_i denote the lifetime / failure time of ith item.

Suppose T_i's be iid having a continuous distribution with pdf $f_\theta(t)$ and cdf $F_\theta(t)$ where θ : parameter of the distribution.

Then given $t_{(1)}, \ldots, t_{(r)}$; the realization of $T_{(1)}, \ldots, T_{(r)}$, the likelihood of θ under Type II censoring is

$$L(\theta) = \frac{n!}{(n-r)!} f_\theta(t_{(1)}) \ldots f_\theta(t_{(r)}) \left[\overline{F}_\theta(t_{(r)})\right]^{n-r}$$

MLE under Survival Data: Type I and II Censoring

Verification:

From theory of order statistics, the jount pdf of all the ordeer statistics $T_{(1)}, \ldots, T_{(n)}$ is

$$h_\theta(t_{(1)}, \ldots, t_{(n)}) = n! \prod_{i=1}^{n} f_\theta(t_{(i)})$$

$\Longrightarrow$ the marginal joint pdf of $T_{(1)}, \ldots, T_{(r)}$ at $t_{(1)}, \ldots, t_{(r)}$ will be

$$g_\theta(t_{(1)}, \ldots, t_{(r)}) = \int \ldots \int n! \prod_{i=1}^{n} f_\theta(t_{(i)}) dt_{(n)} \ldots dt_{(r+1)}$$

$$= n! \prod_{i=1}^{r} f_\theta(t_{(i)}) \int \ldots \left[\int_{t_{(n-1)}}^{\infty} f_\theta(t_{(n)}) dt_{(n)} \right] f_\theta(t_{(n-1)}) \ldots$$

$$f_\theta(t_{(r+1)}) dt_{(n-1)} \ldots dt_{(r+1)}$$

MLE under Survival Data: Type I and II Censoring

$$= n!\prod_{i=1}^{r} f_\theta(t_{(i)}) \int \dots \left[\int_{t_{(n-2)}}^{\infty} (1 - F_\theta(t_{(n-1)}))f_\theta(t_{(n-1)})dt_{(n-1)} \right] f_\theta(t_{(n-2)}) \dots$$

$$= n!\prod_{i=1}^{r} f_\theta(t_{(i)}) \int \dots \left[\int_{t_{(n-3)}}^{\infty} \frac{(1 - F_\theta(t_{(n-2)}))^2}{2} f_\theta(t_{(n-2)})dt_{(n-2)} \right] f_\theta(t_{(n-3)}) \dots$$

$$= \frac{n!}{2 \times 3}\prod_{i=1}^{r} f_\theta(t_{(i)}) \int \dots \left[\int_{t_{(n-4)}}^{\infty} (1 - F_\theta(t_{(n-3)}))^3 f_\theta(t_{(n-3)})dt_{(n-3)} \right]$$

$$f_\theta(t_{(n-4)}) \dots$$

$\Longrightarrow$ finally we get $g_\theta(t_{(1)}, \dots, t_{(r)}) = \frac{n!}{(n-r)!}\prod_{i=1}^{r} f_\theta(t_{(i)})[\overline{F}_\theta(t_{(r)})]^{n-r}$.

MLE under Survival Data: Type I and II Censoring

But given the realizations, the form of joint pdf $\Longrightarrow$ the likelihood of θ. Hence the form of the likelihood is verified.

Here $\overline{F}_\theta(t)$: the survival function at the time point t.

Illustration : Let the of an item be exponential with mean θ.

$\Longrightarrow$ the pdf $f_\theta(t) = \frac{1}{\theta} \exp\{-t/\theta\}$

and the survival function at the time point t is

$$\overline{F}_\theta(t) = \exp\{-t/\theta\}$$

MLE under Survival Data: Type I and II Censoring

$\Longrightarrow$ under Type II censoring, the likelihood function of θ will be

$$L(\theta) = \frac{n!}{(n-r)!}\frac{1}{\theta^r}\exp\left[-\left(\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right)/\theta\right]$$

$$\Longrightarrow l(\theta) = \text{Const } - r\log\theta - \frac{\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}}{\theta}$$

$$\Longrightarrow l'(\theta) = -\frac{r}{\theta} + \frac{\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}}{\theta^2}$$

$\Longrightarrow$ the unique solution of likelihood equation $l'(\theta) = 0$ will be

$$\hat{\theta} = \frac{\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}}{r}$$

MLE under Survival Data: Type I and II Censoring

From SOC we can ensure that $\hat{\theta}$ maximises $L(\theta)$ i.e. $\hat{\theta}$ is the MLE of θ under type II censoring.

Note. $\hat{\theta} = \dfrac{\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}}{r}$ is the MVUE of θ.

Verification. Joint pdf of $T_{(1)}, \ldots, T_{(r)}$ is

$$g_\theta(\mathbf{t}_{(\cdot)}) = \frac{n!}{(n-r)!} \frac{1}{\theta^r} \exp\left\{ -\frac{\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}}{\theta} \right\}$$

Define $Z_1 = nT_{(1)}$, $Z_i = (n-i+1)(T_{(i)} - T_{(i-1)})$; $i = 2, \ldots, r$

MLE under Survival Data: Type I and II Censoring

Check that the Jacobian of transformation from $(T_{(1)}, \ldots, T_{(r)}) \longrightarrow (Z_1, \ldots, Z_r)$ is $\frac{(n-r)!}{n!}$ and

$$\sum_{i=1}^{r} Z_i = \sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)} = r\hat{\theta}.$$

$\Longrightarrow$ the joint pdf of $Z_1, \ldots, Z_r$ is

$$h_\theta(\mathbf{z}) = \frac{1}{\theta^r} \exp\left\{ -\frac{\sum_{i=1}^{r} z_i}{\theta} \right\}$$

which implies that $Z_1, \ldots, Z_r$ are iid exponential (mean $= \theta$) random variables.

MLE under Survival Data: Type I and II Censoring

$$\implies E_\theta(\hat{\theta}) = E_\theta\left(\sum_{i=1}^{r} Z_i\right)/r = \theta.$$

Also $g_\theta(\mathbf{t}_{(\cdot)}) \in$ an OPEF. $\implies$ the statistic $\hat{\theta} = \dfrac{\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}}{r}$ is complete sufficient. Hence by Lehman-Scheffe Theorem the **Note** follows.

Type I Censoring

Sometimes experiments are run over a fixed period of time $\ni$ the exact lifetime of an item will be known only if it is less than some pre-determined value.

MLE under Survival Data: Type I and II Censoring

In such a situation data are said to be type I censored (from right).

More precisely a type I censored sample is one that arises when

- n items numbered say $1, 2, \ldots, n$ are subject to limited periods of observations, and
- let $L_1, \ldots, L_n$ be those periods $\ni$
- ith item's lifetime T_i is observable only if $T_i \leq L_i$.
- L_i : called fixed censoring time for ith item
- If all L_i are equal, data are said to be single type I censored.

MLE under Survival Data: Type I and II Censoring

Assume that T_is are iid with common pdf $f_\theta(t)$ and survival function $\overline{F}_\theta(t)$.

From ith item we record the exact the exact lifetime T_i as the realization provided $T_i \leq L_i$. Otherwise L_i is recorded as the realization.

Let Y_i denote the potential response (the response which is surely obtained) from ith item.

MLE under Survival Data: Type I and II Censoring

Then

$$
\begin{aligned}
Y_i &= T_i \text{ if } T_i \leq L_i \text{ (called uncensored case)} \\
&= L_i \text{ if } T_i > L_i \text{ (called censored case)}
\end{aligned}
$$

for all i. $\implies Y_i = \min\{T_i, L_i\}$.

Also define indicator variables

$$
\begin{aligned}
\delta_i &= 1 \text{ if } T_i \leq L_i \text{ (called uncensored case)} \\
&= 0 \text{ if } T_i > L_i \text{ (called censored case)}
\end{aligned}
$$

Then δ_is are called censoring indicators.

So the type I censored data can be represented by the pairs of random variables (Y_i, δ_i) for all i.

$\Longrightarrow$ the jount likelihood of θ for given data set $\{(t_i, \delta_i), i = 1, \ldots, n\}$ on (Y_i, δ_i)s will be

$$L(\theta) = \prod_{i=1}^{n} [f_\theta(t_i)]^{\delta_i} \left[\overline{F}_\theta(L_i)\right]^{(1-\delta_i)}$$

How this is obtained ? It is true that $P_\theta[Y_i = y_i \mid \delta_i = 0] = 1$ if $y_i = L_i$.

$$\begin{aligned} P_\theta[Y_i = y_i, \delta_i = 0] &= P_\theta[\delta_i = 0] = P_\theta[T_i > L_i] \, \text{if} \, y_i = L_i \\ &= \overline{F}_\theta(L_i) \end{aligned}$$

i.e. the likelihood for the ith item is

$$L_i(\theta) = \overline{F}_\theta(L_i) \, \text{if} \, \delta_i = 0 (\Leftrightarrow \; y_i = L_i) \quad \ldots\ldots \quad (*)$$

MLE under Survival Data: Type I and II Censoring

Again

$$\begin{aligned} P_\theta[Y_i \le y_i, \delta_i = 1] &= P_\theta[T_i \le y_i] \text{ (as } \delta_i = 1 \Leftrightarrow T_i \le L_i \Leftrightarrow Y_i = T_i) \\ &= F_\theta(y_i) \end{aligned}$$

$\implies L_i(\theta) = f_\theta(y_i)$ if $\delta_i = 1 \quad \ldots\ldots \quad (**)$

$(*)$ and $(**)$ $\implies L_i(\theta) = [f_\theta(t_i)]^{\delta_i} \left[\overline{F}_\theta(L_i)\right]^{(1-\delta_i)}$

As pairs (Y_i, δ_i)s are independent, the joint likelihood of θ will be

$$L(\theta) = \prod_{i=1}^{n} [f_\theta(t_i)]^{\delta_i} \left[\overline{F}_\theta(L_i)\right]^{(1-\delta_i)}$$

MLE under Survival Data: Type I and II Censoring

Suppose the readings (in some suitable unit) of life from 10 items, set on an experimentation, are as follows:

1.4^* , 0.17 , 1.4^* , 1.4^* , 0.28 , 0.94 , 1.4^* , 0.7 , 1.07 , 1.20
where reading with * is censored from right. If the life distribution is Weibull with density

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}} , \; t > 0 \; ; \; \alpha, \beta > 0,$$

and also $\alpha = 1$, find the ML estimate of β from the life data readings.

Computation. From the nature of censoring, the data are type I censored from right and has the common censoring time point 1.4.

MLE under Survival Data: Type I and II Censoring

Also here β is the only unknown parameter to be estimated.
Note that for the given Weibull distribution, $\overline{F}(t) = e^{-\alpha t^{\beta}}$. Hence the likelihood function of β (with $\alpha = 1$) will be

$$L(\beta) = \prod_{i=1}^{10} \{f(t_i)\}^{\delta_i} \left\{\overline{F}(L)\right\}^{1-\delta_i}$$

where δ_i : censoring indicator and L : the common censoring time ($= 1.4$ here). Therefore

$$L(\beta) = \beta^{r} \prod_{i=1}^{10} t_i^{(\beta-1)\delta_i} e^{-\sum_{i=1}^{10} \delta_i t_i^{\beta}} e^{-(10-r)L^{\beta}}$$

where $r = \sum_{i=1}^{10} \delta_i =$ number of uncensored cases. t_is denote exact readings.

MLE under Survival Data: Type I and II Censoring

So the loglikelihood of β will be

$$\ell(\beta) = r \ln \beta + (\beta - 1)\sum_{i=1}^{10} \delta_i \ln t_i - \sum_{i=1}^{10} \delta_i t_i^{\beta} - (10 - r)L^{\beta}$$

So $$\frac{\partial}{\partial \beta}\ell(\beta) = \frac{r}{\beta} + \sum_{i=1}^{10} \delta_i \ln t_i - \sum_{i=1}^{10} \delta_i t_i^{\beta} \ln t_i - (10 - r)L^{\beta} \ln L.$$

Hence the likelihood equation of β reduces to the form

$$\beta = r\left[\sum_{i=1}^{10} \delta_i t_i^{\beta} \ln t_i + (10 - r)L^{\beta} \ln L - \sum_{i=1}^{10} \delta_i \ln t_i\right]^{-1} \quad \ldots\ldots (*)$$

MLE under Survival Data: Type I and II Censoring

(∗) does not have any explicit solution. So we have to solve it numerically (using Newton Raphson method) for the ML estimate of β. To find out the initial value of β, we used the quantile method.

R program for the solution of numerical equation (∗) :

R Code and Output :

```
> t=c(1.4,0.17,1.4,1.4,0.28,0.94,1.4,0.7,1.07,1.20)
> del=c(0,1,0,0,1,1,0,1,1,1)
> max.iter=100
> r=sum(del)
> # initial value of 'beta'
> qu=quantile(t)
> init=log(log(4))/log(as.numeric(qu[4]))
> init
[1] 0.9707614
> beta=NULL
> beta[1]=init
```

MLE under Survival Data: Type I and II Censoring

R Code and Output (continued) :

```
> # 1st derivative function
> fun1=function(b)
{
sum1=0
sum2=0
for(i in 1:length(t))
{
sum1=sum1+(del[i]*log(t[i]))
}
for(j in 1:length(t))
{
sum2=sum2+(del[j]*((t[j])^b)*log(t[j]))
}
l1=(r/b)+(sum1)-sum2-((10-r)*((1.4)^b)*log(1.4))
return(l1)
}
```

MLE under Survival Data: Type I and II Censoring

R Code and Output (continued) :

```
> # 2nd derivative function
> fun2=function(b)
{
sum1=0
for(i in 1:length(t))
{
sum1=sum1+(del[i]*((t[i])^b)*log(t[i])*log(t[i]))
}
l2=-(r/(b*b))-sum1-((10-r)*((1.4)^b)*log(1.4)*log(1.4))
return(l2)
}
> for(k in 2:max.iter)
{
beta[k]=beta[k-1]-(fun1(beta[k-1])/fun2(beta[k-1]))
if(beta[k]-beta[k-1]<0.0000001)
break
}
> # MLE of 'beta' (converged value)
> beta[k]
[1] 1.244887
```

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Module 16

Saurav De

Department of Statistics
Presidency University

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

An illustration under type I censoring:

Consider exponential life of an item with mean θ.
Under type I censoring the likelihood of θ will be

$$L(\theta) = \prod_{i=1}^{n} [f_\theta(t_i)]^{\delta_i} \left[\overline{F}_\theta(L_i)\right]^{(1-\delta_i)}$$

where $f_\theta(t) = \frac{1}{\theta} \exp\{-t/\theta\}$ and $\overline{F}_\theta(t) = \exp\{-t/\theta\}$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

If we define a set $\mathcal{A} = \{i \mid \delta_i = 1\}$, we can write

$$\begin{aligned} L(\theta) &= \prod_{i \in \mathcal{A}} \frac{1}{\theta} \exp\{-t_i/\theta\} \,.\, \prod_{i \in \mathcal{A}^c} \exp\{-L_i/\theta\} \\ &= \frac{1}{\theta^r} \exp\left\{-\frac{1}{\theta}\left[\sum_{i \in \mathcal{A}} t_i + \sum_{i \in \mathcal{A}^c} L_i\right]\right\} \end{aligned}$$

where $r = \sum_{i \in \mathcal{A}} \delta_i = \sum_{i=1}^{n} \delta_i = \#$ uncensored cases observed.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Now we can write $\sum_{i \in \mathcal{A}} t_i + \sum_{i \in \mathcal{A}^c} L_i = \sum_{i=1}^{n} y_i$

$$\implies L(\theta) = \frac{1}{\theta^r} \exp\left\{ -\sum_{i=1}^{n} y_i/\theta \right\}$$

Hence get the MLE of θ as $\hat{\theta} = \dfrac{\sum_{i=1}^{n} Y_i}{r}$

Note: Here r is not fixed but a random variable.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

$\Longrightarrow$ It is difficult to find the exact sampling distribution of $\hat{\theta}$.
Therefore consider the asymptotic distribution of $\hat{\theta}$.
Asymptotic distribution of MLE $\Longrightarrow$

$$\hat{\theta} \equiv \hat{\theta}_n \xrightarrow{\mathcal{D}} N\left(\theta, I^{-1}(\theta)\right)$$

where $I(\theta)$ = Fisher's Information

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

i.e. $I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log L(\theta)\right]$

Now $L(\theta) = \frac{1}{\theta^r} \exp\left\{-\sum_{i=1}^{n} y_i/\theta\right\}$

$$\implies \log L(\theta) = -r\log\theta - \frac{\sum_{i=1}^{n} y_i}{\theta} \implies \frac{\partial}{\partial\theta}\log L(\theta) = -\frac{r}{\theta} + \frac{\sum_{i=1}^{n} y_i}{\theta^2}$$

And

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta) = \frac{r}{\theta^2} - 2\frac{\sum_{i=1}^{n} y_i}{\theta^3}$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

$$\implies E\left(-\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right) = \frac{2\sum E(Y_i)}{\theta^3} - \frac{E(r)}{\theta^2} \quad \ldots \;\; \ldots \quad (*)$$

Now

$$E(Y_i) = E(Y_i \mid \delta_i = 1) P_\theta[\delta_i = 1] + E(Y_i \mid \delta_i = 0) P_\theta[\delta_i = 0] \; \ldots \ldots \; (**)$$

As $P_\theta[Y_i = L_i \mid \delta_i = 0] = 1$ so $E(Y_i \mid \delta_i = 0) = L_i$.

Also $E(Y_i \mid \delta_i = 1) E(T_i \mid T_i \leq L_i) = \dfrac{\int_0^{L_i} t\frac{1}{\theta}\exp\{-t/\theta\}dt}{F_\theta(L_i)}$

$$= \theta\left(1 - \exp\{-L_i/\theta\}\right)^{-1}\left[1 - \left(\frac{L_i}{\theta} + 1\right)\exp\{-L_i/\theta\}\right]$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

On simplification $(**) \implies$

$$
\begin{aligned}
E(Y_i) &= \theta\left[1 - \left(\frac{L_i}{\theta} + 1\right)\exp\{-L_i/\theta\}\right] + L_i \exp\{-L_i/\theta\} \\
&= \theta\left(1 - \exp\{-L_i/\theta\}\right)
\end{aligned}
$$

$$
\begin{aligned}
\implies E(r) = \sum_{i=1}^{n} E(\delta_i) &= \sum_{i=1}^{n} P_\theta[\delta_i = 1] \\
&= \sum_{i=1}^{n} \left(1 - \exp\{-L_i/\theta\}\right) = Q \text{ say}
\end{aligned}
$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Hence finally $(*) \implies$

$$I(\theta) = \frac{2\theta \sum_{i=1}^{n} (1 - \exp\{-L_i/\theta\})}{\theta^3} - \frac{Q}{\theta^2} = \frac{Q}{\theta^2}$$

$$\implies \hat{Q}_n \xrightarrow{\mathcal{D}} N\left(0, \frac{\theta^2}{Q}\right)$$

Note: MLE of Q will be $\hat{Q} = \sum_{i=1}^{n} \left(1 - \exp\left\{-L_i/\hat{\theta}\right\}\right)$

(by invariance property of MLE)

epathshala पाठशाला A Gateway to All Post Graduate Courses

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Note: For large n, the normality assumption for $\hat{\theta}$ is good. However for small n, it is rather poor. There are alternative approximate methods that can be recommended in the context of asymptotic normality even for small n. One such approach discussed below is due to D. A. Sprott (1973, Biometrika).
Sprott showed that the transformation $\hat{\phi} = \hat{\theta}^{-1/3}$ converges in distribution to normality more closely than $\hat{\theta}$ itself, even for small n.

Obviously here $\phi = \theta^{-1/3}$. Also fom Taylor's expansion we get

$$E(\hat{\phi}) \approx \phi = \hat{\theta}^{-1/3}$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

$$\text{And } V(\hat{\phi}) \approx \left(\frac{\partial}{\partial\theta}\phi\right)^2 V_{asy}(\hat{\theta})$$
$$= \left(-\frac{1}{3}\theta^{-4/3}\right)^2 \frac{\theta^2}{Q} = \frac{\theta^{-2/3}}{9Q} = \frac{\phi^2}{9Q}$$

Thus for testing any hypothesis or constructing any confidence interval of some parametic function of θ, we can start with

$$\frac{\hat{\phi}-\phi}{\sqrt{\frac{\phi^2}{9Q}}} \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{or,} \qquad \frac{\hat{\phi}-\phi}{\sqrt{\frac{\hat{\phi}^2}{9\hat{Q}}}} \xrightarrow{\mathcal{D}} N(0,1).$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Random Censoring : the most frequently used censoring in statistics.

A random censoring is very similar to type I censoring. The only difference is that here the censoring times are also random variables.

A simple random censoring process is one in which ith item is assumed to have lifetime T_i and censoring time C_i with T_i and C_i independently distributed continuous random variables.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

- Let T_i be iid with common pdf $f_\theta(\cdot)$ and common cdf $F_\theta(\cdot)$
- Let C_i be iid with common pdf $g(\cdot)$ and common cdf $G(\cdot)$.

Define (Y_i, δ_i) same way as in type I censoring.
Then the likelihood function of θ for given data set $\{(t_i, \delta_i), i = 1, \ldots, n\}$ on the paired random variables (Y_i, δ_i)s will be

$$L(\theta) = \prod_{i=1}^{n} \left[f_\theta(t_i)\overline{G}(t_i) \right]^{\delta_i} \left[g(t_i)\overline{F}_\theta(t_i) \right]^{1-\delta_i}$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Verification of the form of likelihood:

Under random censoring we know

$$P_\theta\left[y_i < Y_i \leq y_i + \Delta y_i, \delta_i = 0\right] = P_\theta\left[y_i < C_i \leq y_i + \Delta y_i, T_i > y_i\right]$$

$$(\text{as } \delta_i = 0 \Leftrightarrow \ T_i > C_i \Leftrightarrow \ Y_i = C_i)$$

$$= P\left[y_i < C_i \leq y_i + \Delta y_i\right] P_\theta\left[T_i > y_i\right]$$

$$(\text{as } T_i \text{ and } C_i \text{ are independently distributed})$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

$\implies$ with $\delta_i = 0$ the likelihood corresponding to ith unit will be

$$L_i(\theta) = \lim_{\Delta y_i \to 0} \frac{P_\theta\left[y_i < Y_i \leq y_i + \Delta y_i, \delta_i = 0\right]}{\Delta y_i} = g(y_i)\overline{F}_\theta(y_i)$$

On the other hand

$$P_\theta\left[y_i < Y_i \leq y_i + \Delta y_i, \delta_i = 1\right] = P_\theta\left[y_i < T_i \leq y_i + \Delta y_i, C_i > y_i\right]$$

$$(\text{as now } \delta_i = 1 \Leftrightarrow \ T_i \leq C_i \Leftrightarrow \ Y_i = T_i)$$

$$= P_\theta\left[y_i < T_i \leq y_i + \Delta y_i\right] P\left[C_i > y_i\right]$$

$\implies$ now $L_i(\theta) = f_\theta(y_i)\overline{G}(y_i)$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Thus we have

$$\begin{aligned} L_i(\theta) &= g(t_i)\overline{F}_\theta(t_i) \text{ with } \delta_i = 0 \\ &= f_\theta(t_i)\overline{G}(t_i) \text{ with } \delta_i = 1 \end{aligned}$$

Combining these two forms we can write

$$L_i(\theta) = \left[f_\theta(t_i)\overline{G}(t_i)\right]^{\delta_i} \left[g(t_i)\overline{F}_\theta(t_i)\right]^{1-\delta_i} \text{ for all } i.$$

$\implies$ the joint likelihood function of θ based on given data set on independent (Y_i, δ_i) pairs will be

$$L(\theta) = \prod_{i=1}^{n} \left[f_\theta(t_i)\overline{G}(t_i)\right]^{\delta_i} \left[g(t_i)\overline{F}_\theta(t_i)\right]^{1-\delta_i}$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Informative and Noninformative Random Censoring :

Let θ : Parameter of the probability distribution of lifetime variable, T_i.
If the probability distribution of the censoring variable C_i also involves θ as its parameter, the random censoring is informative.
(**Reason :** Then censoring variables also add information about θ.)

Otherwise, it is noninformative.

e.g. $T_i \sim$ Exponential (θ) , $C_i \sim$ Exponential $(\gamma\theta)$

$$\Longrightarrow \text{ informative censoring.}$$

$T_i \sim$ Exponential (θ) , $C_i \sim$ Exponential (ϕ); ϕ independent of θ

$$\Longrightarrow \text{ noninformative censoring.}$$

Kaplan-Meier (K-M) Estimator/Product Limit (PL) Estimator of survival Function

Let $t_1, \ldots, t_n$ be uncensored sample observations on failure times.

Then a nonparametric estimate of the survival function $S(t)$ at the time point t is given by

$$R_n(t) = \frac{\#observations \geq t}{n} \quad \ldots \ldots \text{ (1)}$$

This is basically the complementary empirical distribution function at the point t.

But usually we cannot expect uncensored failure data due to many practical limitations.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Consider a type I censored sample.

In this case # lifetimes or failure times $\geq t$ may not be known exactly. $\Rightarrow$ we need some modification in (1).

The modified estimator by incorporating appropriate way the sense of type I scensoring , is called the PL estimate of the survival function.

PL estimator $\Longrightarrow$ also known as K-M estimator from the authors who first discussed its properties.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Let there be n items and $k(\leq n)$ distinct failure times $t_1 < t_2 < \ldots < t_k$ observed.
Let $d_j = \#$ failures at time point t_j
In addition to failure times, there are censoring times L_i for items whose lifetimes are not observed due to some reasons.
let $n_j = \#$ items at risk of failing at t_j (i.e. $\#$ items that are functioning and uncensored just prior to t_j).
Then the K-M estimator is defined as

$$\hat{R}_n(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}$$

where $n_{j+1} = n_j - d_j - c_j$; $c_j = \#$ items censored at t_j.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Result : The K-M estimator is a nonparametric MLE of the survival function $S(t)$.

Justification : Let T : Lifetime of a randomly chosen item.
$\implies S(t) = P[T > t]$ at the time point t.
Let $t_1 < t_2 < \ldots < t_k \leq t < t_{k+1}$.
Then we can look upon $S(t)$ as

$$S(t) = P[T > t_1]P[T > t_2 \mid T > t_1] \ldots P[T > t \mid T > t_k]$$

Let $\lambda_j =$ Probability that a randomly chosen item will fail at the time t_j given that it survived at t_{j-1}.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

$\Longrightarrow$

$$\begin{aligned} S(t) &= (1-\lambda_1)(1-\lambda_2)\ldots(1-\lambda_k).1 \\ &= \prod_{j:t_j \leq t} (1-\lambda_j) \end{aligned}$$

As $n_j = \#$ items under the risk of failing at t_j given that they survived at t_{j-1} (like $\#$ independent Bernoulli trials) and

$d_j = \#$ items actually failed at t_j (like random $\#$ successes out of n_j trials)

$\Longrightarrow$ $d_j \sim$ Binomial (n_j, λ_j), for all j

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

We already know that for binomial distribution MLE of λ_j is $\hat{\lambda}_j = \frac{d_j}{n_j}$

$\Longrightarrow$ by invariance property, the MLE of S(t) will be

$$\hat{R}_n(t) = \prod_{j:t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \longrightarrow \text{K-M estimator}$$

Note : d_j always follows binomial distribution irrespective of the parent population distribution F_θ of the lifetime T_i.
$\Longrightarrow$ the distribution of K-M estimator is distribution free i.e. K-M estimator is a nonparametric estimator of the survival function and also the MLE of it under type I censored case.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

The estimated asymptotic variance of the K-M estimator $\hat{R}_n(t)$:

$$\log \hat{R}_n(t) = \sum_{j:t_j \leq t} \log(1 - \hat{\lambda}_j).$$

So, from delta method, the asymptotic variance of $\log \hat{R}_n(t)$ will be

$$\begin{aligned} V\left(\log \hat{R}_n(t)\right) &\approx \sum_{j:t_j \leq t} \left(\frac{\partial}{\partial \hat{\lambda}_j} \log(1 - \hat{\lambda}_j)\right)^2 |_{\hat{\lambda}_j = \lambda_j} V(1 - \hat{\lambda}_j) \\ &= \sum_{j:t_j \leq t} (1 - \lambda_j)^{-2} \frac{\lambda_j(1 - \lambda_j)}{n_j} \\ &= \sum_{j:t_j \leq t} \frac{\lambda_j}{n_j(1 - \lambda_j)} \end{aligned}$$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

So the estimated asymptotic variance of $\log \hat{R}_n(t)$ will be

$$\hat{V}\left(\log \hat{R}_n(t)\right) \approx \sum_{j:t_j \leq t} \frac{\hat{\lambda}_j}{n_j(1-\hat{\lambda}_j)} = \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Again, using delta method we have

$$\begin{aligned}\hat{V}\left(\hat{R}_n(t)\right) &\approx \left(\frac{\partial \hat{R}_n(t)}{\partial \log \hat{R}_n(t)}\right)^2 \hat{V}\left(\log \hat{R}_n(t)\right) \\ &= \left(\hat{R}_n(t)\right)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}\end{aligned}$$

This is known as Greenwood's formula (1926) for asymptotic variance of K-M survival estimator.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Example. Calculate the K-M estimate of $S(t)$ for the following data, where $\delta_i = 1$ if individual i died at time t_i and $\delta_i = 0$ if individual i was censored at that time, for $i = 1, \ldots, 8$.

i	t_i	δ_i
1	2	1
2	5	1
3	8	1
4	11	0
5	12	0
6	15	1
7	20	1
8	23	0

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

The K-M estimates $\hat{R}_n(t_i)$ of the survival function at different time points t_i are given in the following table:

i	t_i	δ_i	n_i	$1-\hat{\lambda}_j$	$\hat{R}_n(t_i)$
1	2	1	8	$\frac{7}{8}$	0.875
2	5	1	7	$\frac{6}{7}$	$\frac{7}{8}\cdot\frac{6}{7} = 0.750$
3	8	1	6	$\frac{5}{6}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6} = 0.625$
4	11	0	5	$\frac{5}{5}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{5}{5} = 0.625$
5	12	0	4	$\frac{4}{4}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{5}{5}\cdot\frac{4}{4} = 0.625$
6	15	1	3	$\frac{2}{3}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{5}{5}\cdot\frac{4}{4}\cdot\frac{2}{3} = 0.417$
7	20	1	2	$\frac{1}{2}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{5}{5}\cdot\frac{4}{4}\cdot\frac{2}{3}\cdot\frac{1}{2} = 0.208$
8	23	0	1	$\frac{1}{1}$	$\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{5}{5}\cdot\frac{4}{4}\cdot\frac{2}{3}\cdot\frac{1}{2}\cdot\frac{1}{1} = 0.208$

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Example Consider the following failure time data measured in some suitable unit :

6 4 4 10 5* 5 8* 11 6 8* 6* 8 4* 4 4 7* 6 10* 3 5

where $*$ denotes the reading is right censored, not exact failure time.

Using R program, find the estimates of the survival function at different time points assuming (i) the failure time $T \sim$ shifted Exponential distribution with density $f(t) = \frac{1}{\sigma}e^{-\frac{(t-\alpha)}{\sigma}}$, $t \geq \alpha > 0$ and (ii) $T \sim F$, where F is absolutely continuous.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Solution. Given data are Type I censored data.

(i) Assuming shifted exponential failure distribution, we get from Module 15 that the ML estimate of $(\alpha, \sigma) = (y_{(1)}, \frac{\sum(y_i - y_{(1)})}{r})$ where y_is are the potential responses, $y_{(1)} = \min\{y_1, \ldots, y_n\}$ and $r = \#$ uncensored cases.

Also the survival function at time point t will be
$S(t) = P[T > t] = e^{-\frac{(t-\alpha)}{\sigma}}$. Hence its (ML) estimate will be
$\hat{S}_n(t) = e^{-\frac{(t-\hat{\alpha})}{\hat{\sigma}}}$. Using R program we will find these estimates.

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

Assumption (ii) leads to the nonparametric estimation of the survival function under Type I censoring. Hence we will use K-M estimator to estimate the survival probability at each time point. The R program in support of this computation and the findings are as follows.

R program for the estimation of survival probabilities :

R Code and Output :

```
> library(survival)
> t=c(6,4,4,10,5,5,8,11,6,8,6,8,4,4,4,7,6,10,3,5)
> del=c(1,1,1,1,0,1,0,1,1,0,0,1,0,1,1,0,1,0,1,1)
> df=data.frame(t,del)
> # parametric
> alpha=min(t)
> sigma=sum(t-alpha)/sum(del)
> Sn1=exp(-(t-alpha)/sigma)
```

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

R Code and Output (continued) :

```
> Sn1
 [1] 0.5436906 0.8161762 0.8161762 0.2412612 0.6661436 0.6661436
     0.3621760
 [8] 0.1969117 0.5436906 0.3621760 0.5436906 0.3621760 0.8161762
     0.8161762
[15] 0.8161762 0.4437473 0.5436906 0.2412612 1.0000000 0.6661436
> # non-parametric
> fit=survfit(Surv(df$t,df$del)~1,df)
```

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

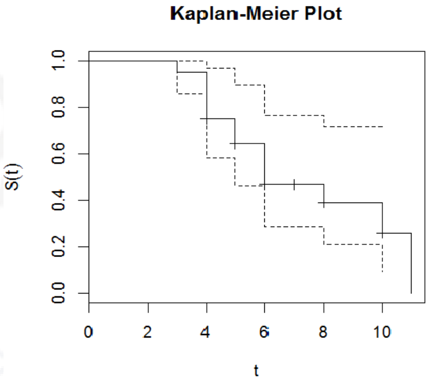
R Code and Output (continued) :

```
> summary(fit)
Call: survfit(formula = Surv(df$t, df$del) ~ 1, data = df)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3     20       1    0.950  0.0487       0.8591        1.000
    4     19       4    0.750  0.0968       0.5823        0.966
    5     14       2    0.643  0.1087       0.4616        0.895
    6     11       3    0.468  0.1170       0.2862        0.764
    8      6       1    0.390  0.1207       0.2123        0.715
   10      3       1    0.260  0.1331       0.0951        0.709
   11      1       1    0.000     NaN           NA           NA
> Sn2=fit$surv
> Sn2
[1] 0.9500000 0.7500000 0.6428571 0.4675325 0.4675325 0.3896104
    0.2597403
[8] 0.0000000
> # K-M plot
> plot(fit,xlab="t",ylab=expression(S(t)),main="Kaplan-Meier Plot")
```

MLE under Survival Data: Type I and Random Censoring and K-M Estimator

The Kaplan-Meier Plot corresponding to the problem :



Partial Likelihood and Cox Proportional Hazard Model

Module 17

Saurav De

Department of Statistics
Presidency University

Partial Likelihood and Cox Proportional Hazard Model

- Let T : lifetime of an item following continuous probability distribution and $P(T \geq 0) = 1$.
- $\overline{F}$(t) or S(t) or $P(T > t)$: Survival probability of the item at time point t.
- S(t) : known as survival function at time point t.
- H(t) $=$ [- log_eS(t)] : cumulative hazard function at time point t.

Partial Likelihood and Cox Proportional Hazard Model

- h(t) = $\frac{d}{dt}$ H(t) = $\frac{d}{dt}$[- log_eS(t)] : instantaneous failure/hazard rate at time point t.
- h(t) : known as hazard function at time point t.

Note : h(t) = $\frac{-\frac{d}{dt}\overline{F}(t)}{\overline{F}(t)} = \frac{f(t)}{\overline{F}(t)}$, where f(t) : lifetime density at time point t.

- S(t) $\downarrow t \Leftrightarrow$ [- log_eS(t)] $\uparrow t$. So, h(t) = $\frac{d}{dt}$ [- log_e S(t)] $\geq 0, \forall\ t$.

Partial Likelihood and Cox Proportional Hazard Model

Survival Regression or Hazard Regression :

- Survival or hazard of an item at time t usually depends on some characteristics like gender, age, Body Mass Index (BMI), etc. Such characteristics are known as covariates.
- Let x denote a covariate supposed to affect $h(t)$, now expressed as $h(t, x)$, without loss of generality.
- Thus, survival or hazard regression can be looked upon as the functional dependence of $h(t, x)$ on covariate x.

Partial Likelihood and Cox Proportional Hazard Model

Purposes of hazard regression :

- To explore a mathematical model like $h(t, x) = g(x)$, where $g(x)$: suitable functional form.
- To predict an estimate of hazard corresponding to some covariate value.
- To measure effect of covariate x on hazard function, $h(t, x)$.
- To test the significance of covariate information x on $h(t, x)$.

Partial Likelihood and Cox Proportional Hazard Model

Cox's Proportional Hazard Model :

$$h(t, \mathbf{x}) = h_0(t, \alpha)\, e^{\beta' \mathbf{x}}$$

, where $\mathbf{x}$ denotes a vector of covariates;
α and β denote the parameters of the Cox's regression model.

Partial Likelihood and Cox Proportional Hazard Model

Some realisations on Cox's regression model :

- $h_0(t, \alpha)$: the value of $h(t, \mathbf{x})$ at $\mathbf{x} = \mathbf{0}$ (i.e., a baseline choice of $\mathbf{x}$). That is why, $h_0(t, \alpha)$ is called baseline hazard function at time point t.
- $h_0(t, \alpha)$ depends on t, but not on covariates.
- Term $e^{\beta' \mathbf{x}}$ depends on covariates $\mathbf{x}$, but not on time point t.
- β measures the effects of $\mathbf{x}$ on $h(t, \mathbf{x})$. How ?
  For univariate $\mathbf{x}$, denoted by x, $\frac{h(t,x+1)}{h(t,x)} = e^{\beta} \Rightarrow \beta$ gauges the impact of change in $h(t, x)$ for unit change in x value.

Some realisations on Cox's regression model (continued) :

- Consider two individuals with covariates $\mathbf{x_1}$ and $\mathbf{x_2}$. Then, the ratio of their hazards at time t is given by :

$$\frac{h(t, \mathbf{x_1})}{h(t, \mathbf{x_2})} = \frac{h_0(t, \alpha)\, e^{\beta' \mathbf{x_1}}}{h_0(t, \alpha)\, e^{\beta' \mathbf{x_2}}} = e^{\beta' (\mathbf{x_1} - \mathbf{x_2})},$$

which is constant with respect to time t

Partial Likelihood and Cox Proportional Hazard Model

$\Rightarrow h(t, \mathbf{x_1}) \propto h(t, \mathbf{x_2})$ with respect to time t.

- In other words, the hazards are proportional $\forall\, t$. Hence, the name *Proportional Hazard Model (P.H.M.)*.

  i.e. hazard ratio of two individuals at age 8 years and age 80 years will remain same. It seems to be little unrealistic at times.

Partial Likelihood and Cox Proportional Hazard Model

Some realisations on Cox's regression model (continued) :

- In Cox's P.H.M., one will be interested in parameters β, which measure the effects of covariates on survival, but usually we are not interested in α.
  $\Rightarrow \beta \rightarrow$ parameters of interest, $\alpha \rightarrow$ nuisance parameters.
- Hence, no form is pre-specified for baseline hazard, $h_0(t, \alpha)$. The Cox's P.H.M. is thus called a semi-parametric model.

Estimating the covariate parameters β : Partial Likelihood

- As here, the form of $h_0(t, \alpha)$ is not specified, so the form of likelihood function is not known properly $\Rightarrow$ the usual method of maximum likelihood fails.
- For estimating the regression parameters $\boldsymbol{\beta}$, Cox developed a non-parametric method and called it as *Partial Likelihood*.

Partial Likelihood and Cox Proportional Hazard Model

- Let m = number of individuals under study.
  $\delta_i = 1$, if i^{th} individual is uncensored.
  $= 0$, if i^{th} individual is right-censored. (i = 1,. . . , m).
- Define R(t_i) = set of all individuals surviving or functioning at time t_i
  $\rightarrow$ Risk set at t_i.
- Let $h_j(t)$ denote hazard of j^{th} individual at time t.
- Then, given that t_i is an event time (i.e., failure/death time), the probability that individual i has that event is given by :

$$P([i]|t_i) = \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)}, \; [i] : individual\, i.$$

Partial Likelihood and Cox Proportional Hazard Model

- As per Cox's P.H.M. assumptions, $h_j(t, \mathbf{x_j}) = h_0(t, \alpha)\, e^{\boldsymbol{\beta}' \mathbf{x_j}}$

  $\Rightarrow P([i]|t_i) = \dfrac{h_0(t_i, \alpha)\, e^{\boldsymbol{\beta}' \mathbf{x_i}}}{\sum_{j \in R(t_i)} h_0(t_i, \alpha)\, e^{\boldsymbol{\beta}' \mathbf{x_j}}} = \dfrac{e^{\boldsymbol{\beta}' \mathbf{x_i}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}' \mathbf{x_j}}} = \dfrac{\phi_i}{\sum_{j \in R(t_i)} \phi_j}, (say)$

  where, $\phi_j = e^{\boldsymbol{\beta}' \mathbf{x_j}}$.
- $P([i]|t_i)$ may be called as Risk Probability of individual i at time point t_i.

Partial Likelihood and Cox Proportional Hazard Model

Partial Likelihood for unique failure times :

- Suppose at each event time only one individual observes the event. Also, the individual observes the events independently. $\Rightarrow$ The partial likelihood for $\boldsymbol{\beta}$ is given by :
$$L_p(\boldsymbol{\beta}, x) = \prod_{i=1}^{m} \left[\frac{\phi_i}{\sum_{j \in R(t_i)} \phi_j} \right]^{\delta_i} = \prod_{i=1}^{m} \left\{ P([i]|t_i) \right\}^{\delta_i}.$$
- Power δ_i means the individuals having the events (failure/death) actually contribute to the likelihood, but not the right-censored cases.

Partial Likelihood and Cox Proportional Hazard Model

Why the name *partial* ? Two reasons for that are :

1. L_p is not the full likelihood function of both the parameters α and β, but β only.
2. L_p does not use the full data; as not the actual times of occurrences are important, but their rankings are, e.g., the likelihood remains same no matter if the individuals i, j and k have event times $\{1,2,3\}$ or $\{10,80,94\}$.

Merit of partial likelihood method : The method involves less assumptions and hence is more robust than full likelihood method.
Demerit of partial likelihood method : The method is less powerful compared to a fully parametric model.

Partial Likelihood and Cox Proportional Hazard Model

Note : Partial likelihood acts exactly in a similar way as the full likelihood.

Let $\widehat{\boldsymbol{\beta}_{\boldsymbol{p}}}$ denote the *Maximum Partial Likelihood* estimate of $\boldsymbol{\beta}$.

$$\Rightarrow L_p(\widehat{\boldsymbol{\beta}_{\boldsymbol{p}}}) = \sup_{\boldsymbol{\beta}_{\boldsymbol{p}}} \prod_{i=1}^{m} \left[\frac{e^{\boldsymbol{\beta}' \mathbf{x_i}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}' \mathbf{x_j}}} \right]^{\delta_i}.$$

Partial Likelihood and Cox Proportional Hazard Model

Properties of $\widehat{\beta_p}$: Maximum Partial Likelihood estimate of β

1. $\widehat{\beta_p} \xrightarrow{P} \beta$ as m $\to \infty$.
2. $Disp(\widehat{\beta_p}) \approx I_p^{-1}$, where I_p is calculated from L_p exactly in the same way as usual *Information Matrix* from full likelihood.
3. $\widehat{\beta_p} \xrightarrow{\mathscr{D}}$ N(β, I_p^{-1}) as m $\to \infty$.

Partial Likelihood and Cox Proportional Hazard Model

Partial Likelihood for repeated failure times :

The case when two or more individuals may have same event times. Let $t_{(i)}$ be i^{th} distinct ordered event time, i.e., $t_{(i)} < t_{(i+1)}$ (that means, ordered) and four failure times 1, 1, 3, 3 $\Rightarrow t_{(1)} = 1$ and $t_{(2)} = 3$ (that means, distinct).

- I = number of unique event times, and
- D(t) be the set of individuals having event at time point t.

Partial Likelihood and Cox Proportional Hazard Model

- Under repeated failure times, there are three popular methods of framing likelihood, which are :
  1. Breslow's method.
  2. Efron's method.
  3. Exact method.
- Consider the following situation first.
  Suppose individuals labelled 1-5 are at risk of failing at $t_{(i)} \Rightarrow R(t_{(i)}) = \{ 1, 2, 3, 4, 5 \}$.
- Let individuals 1-3 actually fail at $t_{(i)} \Rightarrow D(t_{(i)}) = \{ 1, 2, 3 \}$.

Partial Likelihood and Cox Proportional Hazard Model

Then Breslow's method contributed from time $t_{(i)}$ to L_p by :

$$\frac{\phi_1}{(\phi_1+\cdots+\phi_5)} \times \frac{\phi_2}{(\phi_1+\cdots+\phi_5)} \times \frac{\phi_3}{(\phi_1+\cdots+\phi_5)}$$

$\Rightarrow$ formally, we have from Breslow's method that :

$$L_p(\beta, x) = \prod_{i=1}^{I} \left\{ \prod_{j \in D(t_i)} \frac{\phi_j}{\left(\sum_{j \in R(t_j)} \phi_j\right)} \right\} = \prod_{i=1}^{I} \left\{ \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\left(\sum_{j \in R(t_{(i)})} \phi_j\right) \mid D(t_{(i)}) \mid} \right\}$$

where, $\mid D(t_{(i)}) \mid$ = number of individuals in set $D(t_{(i)})$.

It assumes that all the 3 individuals fail simultaneously at $t_{(i)} \rightarrow$ a crude and unrealistic assumption.

Partial Likelihood and Cox Proportional Hazard Model

Due to Efron's method contribution to L_p through illustration is :

$$\frac{\phi_1\,\phi_2\,\phi_3}{\left(\phi_1+\cdots+\phi_5\right)\left(\phi_1+\cdots+\phi_5-\frac{1}{3}(\phi_1+\phi_2+\phi_3)\right)\left(\phi_1+\cdots+\phi_5-\frac{2}{3}(\phi_1+\phi_2+\phi_3)\right)}.$$

It assumes non-instanteinity in failure of 3 individuals and accordingly denominators are adjusted.

As it is not known a priori that who is the first to fail, so one-third of $(\phi_1+\phi_2+\phi_3)$ is adjusted from $\sum_{j=1}^{5}\phi_j$ after one fails. Similarly two-third of $(\phi_1+\phi_2+\phi_3)$ is adjusted after first two individuals fail.

Partial Likelihood and Cox Proportional Hazard Model

So, in general, under Efron's method :

$$L_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\prod_{k=1}^{|D(t_{(i)})|} \left[\sum_{j \in R(t_{(i)})} \phi_j - \frac{k-1}{|D(t_{(i)})|} \sum_{j \in D(t_{(i)})} \phi_j \right]}.$$

This method is much more accurate. Also this method is the default method of partial likelihood for fitting Cox PHM using in R saftware.

Partial Likelihood and Cox Proportional Hazard Model

Using the illustration, finally Exact method contributes :

$$\frac{\phi_1\phi_2\phi_3}{\phi_1\phi_2\phi_3 + \phi_1\phi_2\phi_4 + \phi_1\phi_2\phi_5 + \cdots + \phi_3\phi_4\phi_5}.$$

$\Rightarrow$ in general, Exact method defines :

$$L_p(\beta, x) = \prod_{i=1}^{I} \frac{\sum_{j \in D(t_{(i)})} \phi_j}{\sum_{j \in Q_i} \Phi_q}.$$

where, Q_i = set of all $\mid D(t_{(i)}) \mid$-tuples that could be selected from $R(t_{(i)})$ and Φ_q = the product of ϕ_j's over all members j of a $\mid D(t_{(i)}) \mid$-tuple q.

Note :

- Efron's method is closer to Exact method.
- Under untied case, these three methods are just same with L_p for unique failure time.

Partial likelihood estimate by numerical method:

The above three methods of forming partial likelihood function are quite difficult for the purpose of fitting the (Cox PH) model parameters analytically.
Let β be $k-$dimensional model parameter vector and our objective is to get $\hat{\beta}$ that maximises the log partial likelihood function $\ell_p(\beta)$.
Newton-Raphson method, as discussed below, can be used to get the estimate of parameters :

- Choose an arbitrary value $\beta^{(0)}$ as an initial approximation to β.

Partial Likelihood and Cox Proportional Hazard Model

- Define score vector

$$U(\boldsymbol{\beta}) = \left(\frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \beta_k}\right)'$$

  and the $k \times k$ information matrix $I(\boldsymbol{\beta})$ whose (i,j)th element is $\frac{\partial^2 \ell_p(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}$.

- Get the first approx. $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + I^{-1}(\boldsymbol{\beta}^{(0)})U(\boldsymbol{\beta}^{(0)})$

  Second approx. $\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)} + I^{-1}(\boldsymbol{\beta}^{(1)})U(\boldsymbol{\beta}^{(1)})$ and so on.

- The iterative method will converge at $r+1$th step if $\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\beta}^{(r+1)}$ agree upto certain decimal places, and then ml estimate $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(r)}$ or $\boldsymbol{\beta}^{(r+1)}$.

- Further $\boldsymbol{\beta}^{(0)}$ from $\hat{\boldsymbol{\beta}}$, more is r i.e. less likely is the convergence to $\hat{\boldsymbol{\beta}}$.

Partial Likelihood and Cox Proportional Hazard Model

An Illustration. Let x_i : the body mass index (BMI) of individual i. Suppose a study is conducted on individuals who suffered from heart attack. the following 9 individuals died and t_i : the day(s) to death for the individual i following the attack. Data are:

i	1	2	3	4	5	6	7	8	9
t_i	6	98	189	374	1002	1205	2065	2201	2421
x_i	31.4	21.5	27.1	22.7	35.7	30.7	26.5	28.3	27.9

Partial Likelihood and Cox Proportional Hazard Model

Here Cox PHM : $h(t_i, x_i) = h(t_i, \alpha)e^{\beta x_i}$.
The partial loglikelihood will be

$$\ell_p(\beta) = \beta \sum_{i=1}^{9} x_i - \sum_{i=1}^{9} \log\left(\sum_{j \in R(t_i)} e^{\beta x_j}\right)$$

$$\implies \frac{\partial}{\partial \beta}\ell_p(\beta) = \sum_{i=1}^{9} x_i - \sum_{i=1}^{9} \frac{\sum_{j \in R(t_i)} x_j e^{\beta x_j}}{\sum_{j \in R(t_i)} e^{\beta x_j}} \quad \ldots (*)$$

Partial Likelihood and Cox Proportional Hazard Model

Hence

$$\frac{\partial^2}{\partial\beta^2}\ell_p(\beta) = -\sum_{i=1}^{9} \frac{A}{\left(\sum_{j\in R(t_i)} e^{\beta x_i}\right)^2}, \; (**)$$

where $A = \left(\sum_{j\in R(t_i)} x_j^2 e^{\beta x_i}\right)\left(\sum_{j\in R(t_i)} e^{\beta x_i}\right) - \left(\sum_{j\in R(t_i)} x_j e^{\beta x_i}\right)^2$.

Obviously here $U(\beta) = \frac{\partial}{\partial\beta}\ell_p(\beta)$ and $I(\beta) = -\frac{\partial^2}{\partial\beta^2}\ell_p(\beta)$.

$\Longrightarrow$ the Newton-Raphson recursive update relation is

$$\beta^{(k+1)} = \beta^{(k)} + \frac{U(\beta^{(k)})}{I(\beta^{(k)})}, \; k = 0, 1, \ldots$$

Partial Likelihood and Cox Proportional Hazard Model

Let us choose $\beta^{(0)} = 0$.
Hence from $(*)$ and $(**)$ we compute $U(0) = -2.512$ and $I(0) = 77.13$.
So

$$\beta^{(1)} = 0 - 2.512/77.13 = -0.0326$$

This way we get (on computation)

$$\beta^{(2)} = -0.0326 - 0.069/72.83 = -0.0335 \ ,$$

$$\beta^{(3)} = -0.0335 - 0.000061/72.70 \approx -0.0335$$

As both $\beta^{(2)} \& \beta^{(3)}$ agree upto 4 decimal places $\implies$ the iteration converges correct upto 4 decimal places.
Hence, the ML estimate $\hat{\beta} = -0.0335$(correct upto 4 decimal places).

Fitting Cox PHM with partial likelihood method using R :-

Consider the similar type of data as follows:

i	1	2	3	4	5	6	7	8	9	10
t_i	6	98	98	98	98	189	374	374	374	1002
δ_i	1	1	1	0	1	1	0	1	1	1
x_i	31.4	21.5	23.2	22.5	25	27.1	22.7	22.7	20.6	35.7
z_i	89	93	69	58	84	81	81	74	79	70

Contd . . .

Partial Likelihood and Cox Proportional Hazard Model

i	11	12	13	14	15	16	17	18
t_i	1002	1205	1205	1205	2065	2201	2201	2421
δ_i	1	1	0	1	1	1	0	1
x_i	24.7	30.7	25.1	28.8	26.5	28.3	29.5	27.9
z_i	76	73	67	70	86	71	75	73

where t_i : days to death or right censoring for ith individual,
δ_i : Censorship indicator for individual i; $\delta_i = 1$, if died; $\delta_i = 0$, if censored,
(x_i, z_i) : (BMI , Age) for individual i.

Using R, fit Cox PHM and study different inferential aspects of your fitting based on these repeated-event-time data.

Partial Likelihood and Cox Proportional Hazard Model

Fitting through Breslow's, Efron's and Exact Method of partial likelihood using R:

R Code and Output :

```
> library(survival)
> t=c
   (6,98,98,98,98,189,374,374,374,1002,1002,1205,1205,1205,2065,2201

> delta=c(1,1,1,0,1,1,0,1,1,1,1,1,0,1,1,1,0,1)
> bmi=c
   (31.4,21.5,23.2,22.5,25,27.1,22.7,22.7,20.6,35.7,24.7,30.7,25.1,28

> age=c(89,93,69,58,84,81,81,74,79,70,76,73,67,70,86,71,75,73)
> df=data.frame(t,delta,bmi,age)
> a=coxph(Surv(df$t, df$delta) ~ df$bmi + df$age,df,method = "
   breslow")
```

Partial Likelihood and Cox Proportional Hazard Model

R Code and Output (continued) :

```
> summary(a)
Call:
coxph(formula = Surv(df$t, df$delta) ~ df$bmi + df$age, data = df,
    method = "breslow")

  n= 18, number of events= 14

            coef  exp(coef)  se(coef)      z  Pr(>|z|)
df$bmi  -0.01820    0.98197   0.10412  -0.175   0.8613
df$age   0.07958    1.08283   0.04461   1.784   0.0745 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .
        0.1          1
```

Partial Likelihood and Cox Proportional Hazard Model

R Code and Output (continued) :

```
       exp(coef) exp(-coef) lower .95 upper .95
df$bmi     0.982     1.0184    0.8007     1.204
df$age     1.083     0.9235    0.9922     1.182

Concordance= 0.708  (se = 0.105 )
Rsquare= 0.196   (max possible= 0.963 )
Likelihood ratio test= 3.92  on 2 df,   p=0.1407
Wald test            = 4.1  on 2 df,   p=0.1284
Score (logrank) test = 4.31  on 2 df,   p=0.1158
> b=coxph(Surv(df$t, df$delta) ~ df$bmi + df$age,df,method = "efron
    ")
> summary(b)
Call:
coxph(formula = Surv(df$t, df$delta) ~ df$bmi + df$age, data = df,
    method = "efron")

  n= 18, number of events= 14
```

Partial Likelihood and Cox Proportional Hazard Model

R Code and Output (continued) :

```
             coef  exp(coef)  se(coef)       z  Pr(>|z|)
df$bmi   -0.02690    0.97346   0.10567  -0.255    0.7991
df$age    0.08192    1.08537   0.04461   1.836    0.0663 .
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .
        0.1     1

        exp(coef) exp(-coef) lower .95 upper .95
df$bmi     0.9735     1.0273    0.7913     1.197
df$age     1.0854     0.9213    0.9945     1.185

Concordance= 0.692  (se = 0.105 )
Rsquare= 0.209   (max possible= 0.961 )
Likelihood ratio test= 4.22  on 2 df,   p=0.1211
Wald test            = 4.42  on 2 df,   p=0.11
Score (logrank) test = 4.66  on 2 df,   p=0.0975
```

Partial Likelihood and Cox Proportional Hazard Model

R Code and Output (continued) :

```
> d=coxph(Surv(df$t, df$delta) ~ df$bmi + df$age,df,method = "exact
    ")
> summary(d)
Call:
coxph(formula = Surv(df$t, df$delta) ~ df$bmi + df$age, data = df,
    method = "exact")

  n= 18, number of events= 14

           coef  exp(coef)  se(coef)       z  Pr(>|z|)
df$bmi  -0.02102   0.97920   0.10851  -0.194   0.8464
df$age   0.09451   1.09912   0.05071   1.864   0.0624 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
        0.1          1
```

Partial Likelihood and Cox Proportional Hazard Model

R Code and Output (continued) :

```
        exp(coef) exp(-coef) lower .95 upper .95
df$bmi     0.9792     1.0212    0.7916     1.211
df$age     1.0991     0.9098    0.9951     1.214

Rsquare= 0.221   (max possible= 0.94 )
Likelihood ratio test= 4.5  on 2 df,   p=0.1052
Wald test            = 4.3  on 2 df,   p=0.1162
Score (logrank) test = 4.75  on 2 df,   p=0.09296
```

Conditional and Marginal Likelihood

Module 18

Saurav De

Department of Statistics
Presidency University

Conditional and Marginal Likelihood

- Multiparameter case (i.e., number of parameters ≥ 2) in probability distribution is very common, most realistic too.
- Those cases the full likelihood will be a function of all those parameters.
- But only few of those parameters become the parameters of our interest in true sense with respect to the inferential study. The rests are considered as nuisance parameters. So, we wish to eliminate them while estimating the parameters of interest.

Conditional and marginal likelihoods are two popular alternative likelihood approaches where :

- nuisance parameters can be eliminated.
- likelihood function can be framed solely as a function of parameters of interest.

These approaches are useful as the conventional likelihood (known also as the *full likelihood*) can become unrealistic or fail completely when the setup is overburdened with many or high-dimensional nuisance parameters.

Conditional Likelihood Method :

Let $f(\mathbf{x};\theta,\lambda)$: the joint density of $k(>2)$ random variables $(x_1,\ldots,x_k)=\mathbf{x}$.

Let θ be the real or vector valued parameter of interest and λ be the nuisance parameter (or parameter vector).

If possible, suppose the joint density can be expressed as :

$$f(\mathbf{x};\theta,\lambda)=f(x_1,\ldots,x_r|x_{r+1},\ldots,x_k;\theta)\,f(x_{r+1},\ldots,x_k;\theta,\lambda),$$

i.e., joint density $=$ conditional density $\times$ marginal density,

where $(x_1,\ldots,x_r|x_{r+1},\ldots,x_k)$ is so divided that conditional density becomes free from nuisance parameter λ.

Conditional and Marginal Likelihood

Based on a sample of size n on $\mathbf{x}$, we have :

$$\prod_{i=1}^{n} f(\mathbf{x}_i;\theta,\lambda) = \prod_{i=1}^{n} f(x_{1i},\ldots,x_{ri}|x_{(r+1)i},\ldots,x_{ki};\theta) \prod_{i=1}^{n} f(x_{(r+1)i},\ldots,x_{ki};\theta,\lambda).$$

Obviously the LHS is nothing but the conventional or full likelihood of all the parameters (θ,λ) based on full data, where as the first product in the RHS is called the conditional likelihood of θ given X $(=$ Data$)$ on $x_{r+1},\ldots,x_k$ and denote it by $L(\theta|X)$ for simplicity.

In conditional likelihood method of estimation, we exclusively focus on this $L(\theta|X)$ and nowhere else, though the second product in RHS may also carry information about θ.

Thus the conditional likelihood method may incur some loss of information about the parameter of interest θ, provided θ (or some components of θ in vector-valued case) are common in the conditional and marginal densities.

Still very often, we focus on the inference about θ from its conditional likelihood or log-likelihood, e.g., the focus in econometric models like CLRM is typically on the parameters like $\theta = (\beta, \sigma^2)$ of the conditional density.

Note that in this method we typically choose not to specify the form of the marginal density and only maximise the conditional likelihood to obtain the conditional maximum likelihood estimators of θ.

Conditional Maximum Likelihood Estimator :

In this case considered above, conditional log-likelihood is :

$$\log L(\theta|X) = \sum_{i=1}^{n} \log f(x_{1i}, \ldots, x_{ri} | x_{(r+1)i}, \ldots, x_{ki}; \theta).$$

Conditional and Marginal Likelihood

Without loss of generality assume that θ is an s-component vector, $s \geq 1$. Then the conditional MLE is that value of θ which maximises $\log L(\theta|X)$ or in most cases solves the system of first order conditions (f.o.c.)

$$\frac{\partial \log L(\theta|X)}{\partial \theta} = \mathbf{0},\ i.e.,\ \left(\frac{\partial \log L(\theta|X)}{\partial \theta_1}, \ldots, \frac{\partial \log L(\theta|X)}{\partial \theta_s}\right)' = \mathbf{0}.$$

i.e., the score vector is null where $\frac{\partial \log L(\theta)}{\partial \theta}$ denotes the score function.

The standard regularity conditions :

Let the true conditional density be denoted by $f(\mathbf{x}^1 \mid \mathbf{x}^2, \theta_0)$ where, $\mathbf{x}^1 = (x_1, \ldots, x_r)'$, $\mathbf{x}^2 = (x_{r+1}, \ldots, x_k)'$ and θ_0, the true parameter vector.

Then the regularity conditions used for conditional MLE are

$$\mathbf{R1} : E_{\theta_0}\left[\frac{\partial \log f(\mathbf{x}^1 \mid \mathbf{x}^2; \theta_0)}{\partial \theta}\right] = \mathbf{0},$$

or in other words, the expected value of the score vector is null, expectation being taken with respect to the true conditional density.

R2 : $E_{\theta_0}\left[\frac{\partial^2 \log f(\mathbf{x}^1 \mid \mathbf{x}^2;\theta_0)}{\partial\theta\,\partial\theta'}\right] = E_{\theta_0}\left[\frac{\partial \log f(\mathbf{x}^1 \mid \mathbf{x}^2;\theta_0)}{\partial\,\theta}\cdot\frac{\partial \log f(\mathbf{x}^1 \mid \mathbf{x}^2;\theta_0)}{\partial\,\theta'}\right].$

Now **R2** implies :

$$E_{\theta_0}\left[\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'}\right] = E_{\theta_0}\left[\frac{\partial \log L(\theta_0|X)}{\partial\,\theta}\cdot\frac{\partial \log L(\theta_0|X)}{\partial\,\theta'}\right],$$

which are nothing but the two forms of getting $(s \times s)$ conditional Fisher's Information matrix.

In fact, the true variance matrix of the score vector is :

$$V\left(\frac{\partial \log L(\theta_0|X)}{\partial\theta}\right) = E\left[\frac{\partial \log L(\theta_0|X)}{\partial\,\theta}\cdot\frac{\partial \log L(\theta_0|X)}{\partial\,\theta'}\right]$$

i.e., the $(s \times s)$ conditional Information Matrix.

Now, we consider the asymptotic properties of conditional MLE. Assuming :

1. the true conditional density $f(\mathbf{x}^1 \mid \mathbf{x}^2; \theta)$ is used to define the conditional likelihood function.
2. $f(\mathbf{x}^1 \mid \mathbf{x}^2; \theta_1) = f(\mathbf{x}^1 \mid \mathbf{x}^2; \theta_2)$ if and only if $\theta_1 = \theta_2$ (so the $L(\theta|X)$ has a unique maximum).
3. $\frac{1}{n}\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log L_i(\theta_0|\mathbf{x_i}^2)}{\partial\theta\,\partial\theta'} \xrightarrow{P} A_0,$
   where $L_i(\theta_0|\mathbf{x_i}^2) = f(\mathbf{x_i}^1 \mid \mathbf{x_i}^2; \theta_0)$ and A_0 is an $(s \times s)$ non-singular matrix.
4. Regularity conditions **R1** and **R2** hold.

Conditional and Marginal Likelihood

We have $\hat{\theta}_{CML} \xrightarrow{P} \theta_0$ (weak consistency) ;
$\hat{\theta}_{CML}$ denoting conditional MLE and

$$\sqrt{n}(\hat{\theta}_{CML} - \theta_0) \xrightarrow{\mathscr{D}} N_s(\mathbf{0}, -A_0^{-1}) \text{ (asymptotic normality).}$$

Now we know from assumption **3** that, $A_0 = E_{\theta_0}\left(\frac{\partial^2 \log L_i(\theta_0|\mathbf{x_i}^2)}{\partial\theta\,\partial\theta'}\right)$
$\forall\, i = 1, \ldots, n$ (which is evident from Khinchinte's WLLN).

So, $E_{\theta_0}\left[\frac{1}{n}\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'}\right] = A_0$ or $E_{\theta_0}\left[\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'}\right] = n\,A_0$.

So the asymptotic distribution can also be expressed as :

$$(\hat{\theta}_{CML} - \theta_0) \xrightarrow{\mathscr{D}} N_s(\mathbf{0}, -(n\,A_0)^{-1}) \equiv N_s\left(\mathbf{0}, -\left(E_{\theta_0}\left[\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'}\right]\right)^{-1}\right),$$

where $\left(E_{\theta_0}\left[\frac{\partial^2 \log L(\theta_0|X)}{\partial\theta\,\partial\theta'}\right]\right)^{-1}$: the inverse of the conditional Information Matrix.

To ensure that the conditional density will be a function of θ only, we can use the following mechanism.

For simplicity, we take $k = 1$, i.e., univariate case. Consider a random sample $X_1, \ldots, X_n$ of size n on univariate $X \Rightarrow$ the joint density is now $f(x_1, \ldots, x_n; \theta, \lambda) = \prod_{i=1}^{n} f(x_i; \theta, \lambda)$.

Step 1 : Get the complete sufficient statistic T_λ of the nuisance parameter λ, $T_\lambda = T(X_1, \ldots, X_n) \Rightarrow$ the joint pdf $f(\mathbf{x}; \theta, \lambda)$ can be expressed as :

$$f(\mathbf{x}; \theta, \lambda) = f(\mathbf{x}; \theta | T_\lambda = t).f_{T_\lambda}(t; \theta, \lambda) \ldots\ldots\ldots (*)$$

because

$$f(\mathbf{x}; \theta | T_\lambda = t) = \frac{f(\mathbf{x}; \theta, \lambda)}{f_{T_\lambda}(t; \theta, \lambda)} .$$

Note : T_λ is sufficient for $\lambda \Rightarrow f(\mathbf{x};\theta|T_\lambda)$ is independent of λ.

Step 2 : Given the data $\{x_1,\ldots,x_n\}$ on the random sample $\{X_1,\ldots,X_n\}$, from (*) we get :

$$L(\theta;\lambda|\mathbf{x}) = L(\theta|T_\lambda = t)\,.\,L^*(\theta,\lambda;t).$$

where LHS = full likelihood of θ, λ

and $L(\theta|T_\lambda = t) = \prod_{i=1}^{n} f(x_i;\theta|T_\lambda = t) \rightarrow$ conditional likelihood of θ, given $T_\lambda = t$.

$\ell(\theta|T_\lambda = t) = \log L(\theta|T_\lambda = t) =$ conditional log-likelihood of θ.

Conditional and Marginal Likelihood

In this context, two cases may occur.

- *Case 1 :* For fixed θ_0, $T_\lambda(\theta_0)$ depends on θ_0.
- *Case 2 :* $T_\lambda(\theta_0) = T_\lambda$, i.e., independent of θ_0, $\forall\ \theta_0$.

Following illustrations are useful to describe these two cases :

Illustration 1 : Let $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$.
Let $\sigma^2 \to$ parameter of interest $\Rightarrow \mu \to$ the nuisance parameter.

Now, joint pdf is :

$$f(\mathbf{x}; \sigma^2, \mu) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Conditional and Marginal Likelihood

If $T_\sigma = \sum_{i=1}^n X_i^2$ and $T_\mu = \sum_{i=1}^n X_i$, then we know that (T_μ, T_σ^2) is complete sufficient for $(\mu, \sigma^2) \Rightarrow$ given $\sigma^2 = \sigma_0^2$, $T_\mu = \sum_{i=1}^n X_i$ is sufficient for μ.

Also, T_μ does not depend on σ_0^2, i.e., independent of σ^2, $\forall\ \sigma \Rightarrow$ the present illustration is in favour of *Case 2*.
Now,

$$f(\mathbf{x}; \sigma^2 | T_\mu = t) = \frac{f(\mathbf{x}; \sigma^2, \mu)}{f_{T_\mu}(t)}$$

where, $t = T_\mu(\mathbf{x}) = \sum_{i=1}^n x_i = n\bar{x}$.

Conditional and Marginal Likelihood

It can be shown that on simplification we finally get :

$$f(\mathbf{x}; \sigma^2 | T_\mu = t) = C(\sigma^2)\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$= L(\sigma^2 | T_\mu = t) \rightarrow$ conditional likelihood of σ^2 given $T_\mu = t = n\overline{x}$.

or, $l(\sigma^2 | T_\mu = t)$ (i.e., the conditional log-likelihood)

$$= log_e\, C(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{2\sigma^2} \rightarrow \text{(dependent on } \sigma^2 \text{ only)}$$

$\Rightarrow$ we can conduct statistical inference for σ^2 based on the above conditional log-likelihood.

Now consider the next illustration.

Illustration 2 : Let $X_1 \sim N(\mu_1, 1)$, $X_2 \sim N(\mu_2, 1)$ independently.

Let $\theta = \frac{\mu_2}{\mu_1} \rightarrow$ parameter of interest and $\lambda = \mu_2$ is the nuisance parameter, i.e., $\mu_1 = \frac{\lambda}{\theta}$.

Now, it can be derived that for fixed $\theta = \theta_0$, the sufficient statistic for λ is $T_\lambda(\theta_0) = X_1 + \theta_0 \, X_2$ which follows $N(\mu_1 + \theta_0 \, \mu_2, 1 + \theta_0^2)$.

Here, the sufficient statistic T_λ and its distribution depend on θ_0. So this illustration indicates *Case 1*. It can be shown that the conditional log-likelihood $l(\theta, \theta_0 | T_\lambda(\theta_0) = t)$ is independent of λ.

$\Rightarrow$ we can make statistical inference for θ based on this conditional log-likelihood function, e.g., to find the conditional ML estimate, we can solve the equation

$$\frac{\partial \ell(\theta, \theta_0 | T_\lambda(\theta_0) = t)}{\partial \theta}\Big|_{\theta_0 = \theta} = 0.$$

$\Rightarrow$ on simplification, $\hat{\theta} = \frac{x_2}{x_1}$.

General Approach :

- When $T_\lambda(\theta_0) = T_\lambda$ is independent of θ_0, the conditional log-likelihood (which now depends only on θ) will be

$$\ell_{cond}(\theta) = \log f_\theta(\mathbf{x}|\, T_\lambda) = \log f_{\theta,\lambda}(\mathbf{x}) - \log f_{T_\lambda}(t)$$

  where $f_{T_\lambda}(t)$ denotes the density of T_λ at t.

  Any $\theta = \hat{\theta}_{CML}$ that maximises $\ell_{cond}(\theta)$, is a conditional ML estimate of θ.

  The asymptotic variance (or dispersion in vector valued case) of $\hat{\theta}_{CML}$ will be the inverse of the conditional Fisher's Information $I_{cond}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{cond}(\theta)\right]$.
  Note that both $\hat{\theta}_{CML}$ and $I_{cond}(\theta)$, in general differ from those derived from the full likelihood.

- When $T_\lambda(\theta_0)$ depends on θ_0, the conditional log-likelihood (which now depends on $\theta, \theta_0, \lambda$) will be

$$\ell_{cond}(\theta, \theta_0, \lambda) = \log f_\theta(\mathbf{x} | T_\lambda(\theta_0)) = \log f_{\theta,\lambda}(\mathbf{x}) - \log f_{T_\lambda(\theta_0)}(t)$$

where $f_{T_\lambda(\theta_0)}(t)$ denotes the density of $T_\lambda(\theta_0)$ at t.

Now $\hat{\theta}_{CML}$ is a solution of $\left[\frac{\partial}{\partial\theta}\ell_{cond}(\theta, \theta_0, \lambda)\right]_{\theta_0=\theta, \lambda=\hat{\lambda}(\theta)} = \mathbf{0}$.

The asymptotic variance (or dispersion in vector valued case) of $\hat{\theta}_{CML}$ will be the inverse of the conditional Fisher's Information $I_{cond}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{cond}(\theta, \theta_0, \lambda)\right]_{\theta_0=\theta, \lambda=\hat{\lambda}(\theta)}$.

Conditional and Marginal Likelihood

As a brief outline of marginal likelihood
(also known as integrated likelihood), we can say that in the context of multiparameter, this likelihood is obtained by marginalizing or averaging out some nuisance parameters from the full likelihood model and retaining the parameter(s) of our interest.

In reference with Bayesian statistics, this may be called as the evidence or model evidence. Maintaining the previous notational implication, let the nuisance parameter $\lambda \sim$ density $p(\lambda;\theta)$

$$\Rightarrow \; f(x;\theta) = \int_{\lambda} f(x;\theta,\lambda)\, p(\lambda;\theta)\, d\lambda$$

.

Based on a sample $(x_1,\ldots,x_n) = \mathbf{x}$, the marginal likelihood of θ will be : $L(\theta;\mathbf{x}) = \int_{\lambda} f(\mathbf{x};\theta,\lambda)\, p(\lambda;\theta)\, d\lambda$.

Note : Marginal likelihoods are difficult to compute and the particular solutions are available for a small class of distributions, otherwise numerical methods are applicable.

Profile Likelihood and Associated Confidence Interval

Saurav De

Department of Statistics
Presidency University

Profile Likelihood and Associated Confidence Interval

- For the probability models having more than one unknown parameter, we often concentrate in the inference of one parameter of interest or of a real-valued function of the parameters.
- For illustration, let θ be the parameter of interest and δ be the (vector of) other parameter(s).

Profile Likelihood and Associated Confidence Interval

Let $L(\theta, \delta)$ be the likelihood function.
Then, profile likelihood function of θ, denoted by $L_1(\theta)$ is $= max_\delta\, L(\theta, \delta)$ for each value of θ, i.e., profile likelihood function of θ obtained by maximising out the remaining parameters(s) δ for each fixed value of the parameter of interest.

Profile Likelihood and Associated Confidence Interval

It provides :

- a graphical summary of the most plausible value of θ.
- parameterization invariant procedure.
- corrected confidence intervals that are more accurate than Wald Type confidence intervals.

Profile Likelihood and Associated Confidence Interval

Note : Wald Type confidence interval (CI) may not perform satisfactorily if either :

- distribution of the parameter estimator is notably skewed, or
- standard error is a poor estimate of the standard deviation of the estimator.

Now, asymptotic standard errors under GLMs are obtained from the information matrix.

$\Rightarrow$ Wald's CI may work poorly for small to moderate sample sizes.

Profile Likelihood and Associated Confidence Interval

On the other hand, confidence intervals based on profile likelihood don't need normality assumptions of the estimator, and perform better than Wald Type CI for small sample cases.

Anyway, the asymptotic null distribution of the profile log-likelihood ratio test statistic is chi-square.

Profile Likelihood and Associated Confidence Interval

In general, consider a statistical model with p-dimensional parameter and log-likelihood $l(\theta)$. Let the parameter function $\psi = \psi(\theta)$ be the parameter of interest for inferential work.

Let $\hat{\theta}$ denote the overall MLE of θ and let $\hat{\psi} = \psi(\hat{\theta})$.

Further, let $\hat{\theta}(\psi)$ be the MLE of θ for fixed ψ.
$\Rightarrow$ profile log-likelihood of ψ; $l_p(\psi) = l(\hat{\theta}(\psi))$, $\forall\ \psi$.

Profile Likelihood and Associated Confidence Interval

Interesting features of l_p :

- it is invariant under one-to-one reparameterization of θ, and
- it is defined for any log-likelihood $l(\theta)$; no special structure like exponential family, location-scale model, etc. is needed for its definition.

But, computation of $l_p(\psi)$ requires a constrained maximization for each fixed value of $\psi \Rightarrow$ Direct computation of $l_p(\psi)$ is difficult.

Profile Likelihood and Associated Confidence Interval

Computation of the profile likelihood :

The curve $\hat{\theta}(\psi)$ is computed on a grid and then the log-likelihood $l(\theta)$ is evaluated along that curve. In this context, the strategy for computation of $\hat{\theta}(\psi)$ is as follows :

1. Start at $\hat{\theta}$.
2. Take a small step in the direction of the tangent.
3. Apply Newton-Raphson iterations to move back onto the curve and hence find out a point on that curve.

(continued onto next slide......)

Profile Likelihood and Associated Confidence Interval

4. Repeat steps **1** and **2** such that one moves along the curve $\hat{\theta}(\psi)$ and find a set of points exactly lying on that curve.
5. *Stopping rule* : We stop against a sufficiently large amount of decrease in the profile log-likelihood.
6. Next, we go back to $\hat{\theta}$ and move to the opposite direction.

Let λ be the Lagrangian multiplier for maximising $l(\theta)$ subject to $\psi(\theta) = \psi$.

Profile Likelihood and Associated Confidence Interval

Stein's Least Favourable Family :

Define a function, $g(\theta) = l(\theta) + \lambda(\psi(\theta) - \psi)$ of p-dimensional parameter θ, where λ is the Lagrangian multiplier for maximising $l(\theta)$ subject to the constraint $\psi(\theta) = \psi$.

Let $\triangledown g(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, \ldots, \frac{\partial g(\theta)}{\partial \theta_p})' =$ gradient of g.

Then, $\hat{\theta}(\psi)$ is the solution to $\triangledown g(\theta) = 0$ and $\lambda(\psi)$ be the solution for λ.

$\Rightarrow$ at $\psi = \hat{\psi}$, $\lambda(\psi) = 0$ and the derivative of $\hat{\theta}(\psi) \propto \{\nabla^2 l(\theta)\}^{-1} \nabla \psi(\theta)$ evaluated at $\hat{\theta}$.

Then the line through $\hat{\theta}$ having the said direction is known as *Stein's Least Favourable Family*.
The family is used in place of $\hat{\theta}(\psi)$ for an approximation to the profile log-likelihood.

Profile Likelihood and Associated Confidence Interval

Some simple applications of profile likelihood :

Let $X_1, \ldots, X_n$ be i.i.d. with common pdf :

$$f_{\alpha,\theta}(x) = \frac{1}{\theta} e^{-\frac{x-\alpha}{\theta}}, \; x \geq \alpha.$$
$$= \quad 0, \; otherwise.$$

$$Here, \; 0 < \theta < \infty, \; -\infty < \alpha < \infty.$$

$\Rightarrow$ for each fixed θ, the profile log-likelihood :

$$l_p(\theta) = \max_{\alpha} l(\alpha, \theta), \quad l(\alpha,\theta) : \mathrm{full\,log-likelihood\ of}\ (\alpha,\theta).$$

$$= \quad \max_{\alpha} \left\{ -n \, log_e \theta - \sum_{i=1}^{n} \frac{(x_i - \alpha)}{\theta} \right\}, \quad x_{(1)} \geq \alpha$$

$$= \quad -n \, log_e \theta - \sum_{i=1}^{n} \frac{(x_i - x_{(1)})}{\theta}, \; \forall \; \theta > 0.$$

Profile Likelihood and Associated Confidence Interval

Next,

$$\frac{d\ l_p(\theta)}{d\theta} = 0 \ \Rightarrow \ -\frac{n}{\theta} + \sum_{i=1}^{n} \frac{\left(x_i - x_{(1)}\right)}{\theta^2} = 0 \ .$$

Also,

$$\left.\frac{d^2\ l_p(\theta)}{d\theta^2}\right|_{\hat{\theta}} = \frac{n}{\theta^2} - 2\sum_{i=1}^{n} \frac{\left(x_i - x_{(1)}\right)}{\theta^3} = \frac{n}{\hat{\theta}^2} < 0 \ .$$

$\Rightarrow \hat{\theta}$ is the maximum profile likelihood estimate of θ.

Note that, here θ is the parameter of interest.

Profile Likelihood and Associated Confidence Interval

Similarly, for $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$, the profile log - likelihood of σ^2 will be :

$$l_p(\sigma^2) = \max_{\mu} l(\mu, \sigma^2) = \max_{\mu} \left\{ const. - \frac{n}{2} log_e \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \right\}, \ \forall \ \sigma^2 > 0$$

$$= const. - \frac{n}{2} log_e \sigma^2 - \frac{n}{2} \frac{s^2}{\sigma^2}, \ \forall \ \sigma^2$$

where, $s^2 =$ sample variance with divisor n.

Using maxima-minima principle, we can show that the maximum profile likelihood estimate of σ^2 is s^2 which corresponds to the usual MLE of σ^2.

Profile Likelihood and Associated Confidence Interval

Profile Likelihood Confidence Interval :

Idea of a profile likelihood confidence interval (CI) is to invert a likelihood ratio test to obtain a CI for the parameter of interest.
A simple approach for obtaining CIs from $l_p(\psi)$ is to form the set of values ψ such that $2[l_p(\hat{\psi}) - l_p(\psi)] > C$, i.e., $\{\psi : 2[l_p(\hat{\psi}) - l_p(\psi)] > C\}$.

For $100(1-\alpha)\%$ CI of ψ, choice of C equals to $\chi^2_{\frac{\alpha}{2},1}$, the upper $100\frac{\alpha}{2}\%$ point of χ^2 distribution with 1 degrees of freedom (d.f.). This is because the asymptotic distribution of the LR-statistic is here χ^2_1.

Profile Likelihood and Associated Confidence Interval

The above choice of C produces a CI with an overall coverage $1-\alpha+O(\frac{1}{\sqrt{n}})$.

To achieve higher accuracy in each tail, we consider the signed square root of the LR statistic, i.e., $R(\hat{\psi}) = sgn(\hat{\psi}-\psi)\sqrt{2[l_p(\hat{\psi})-l_p(\psi)]}$.

Then the set $\{\psi : R(\psi) > \sqrt{C}\}$ is identical to the set $\{\psi : 2[l_p(\hat{\psi})-l_p(\psi)] > C\}$. Obviously, for $100(1-\alpha)\%$ CI of ψ, asymptotically $\sqrt{C}$ is chosen as $\tau_{\frac{\alpha}{2}}$; upper $100\frac{\alpha}{2}\%$ point of $N(0,1)$ distribution.

Profile Likelihood and Associated Confidence Interval

Note : With $R(\psi)$ also, the coverage error is $O(\frac{1}{\sqrt{n}})$. But the advantage is that we can have CIs of higher accuracy in each tail by correcting $R(\psi)$ to make its distribution closer to $N(0,1)$.

McCullah (1984) and DiCiccio (1984) have shown that under one-parameter model, in some sense, CIs based on $R(\psi)$ is as good as those obtained from approximate pivots
$\Rightarrow$ $R(\psi)$ might be looked upon as the 'jackknife' of parametric models, as it is simple and can be applied automatically.

Profile Likelihood and Associated Confidence Interval

Logit Model Example :

Let $y_i \sim Bin(n_i, \pi_i),\ i = 1, \ldots, k.$

We fit a logit model :

$$log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \ldots\ldots\ldots (*)$$

where $x_i \rightarrow i^{th}$ value of a single covariate x.

$\Rightarrow$ the log-likelihood function is :

$$\begin{aligned} logL(\beta_0, \beta_1) &= \sum_{i=1}^{k}\left[y_i\, log_e\left(\frac{\pi_i}{1-\pi_i}\right) + n_i log_e(1-\pi_i) + log_e\binom{n_i}{y_i}\right] \\ &= \sum_{i=1}^{k}\left[log_e\binom{n_i}{y_i} + y_i(\beta_0 + \beta_1 x_i) - n_i\, log_e(1 + e^{\beta_0 + \beta_1 x_i})\right]. \end{aligned}$$

Profile Likelihood and Associated Confidence Interval

Naturally, β_1 is the parameter of interest and β_0 the nuisance parameter.
$\Rightarrow$ the profile log-likelihood of β is :

$$l_p(\beta_1) = \max_{\beta_0} log_e \, L(\beta_0, \beta_1), \ \forall \ \beta_1.$$

Now manually it is very difficult and time consuming to get $l_p(\beta_1)$ as well as maximization of $l_p(\beta_1)$ to get maximum profile likelihood estimate of β_1.

The R function $confint(m)$ where m is the fitted logit model, directly provides the profile likelihood CI of β_1.

Quasi-Likelihood Method of Estimation

Module 20

Saurav De

Department of Statistics
Presidency University

In order to construct a likelihood function the probability distribution(s) of the random sample observations should be specified.

But unfortunately(!) often those probability distributions are not available except few information revealing few aspects of the data, like

- the domain of the sample responses;

Continued ...

- how the mean or median response is affected by external stimuli or covariates;
- how variability of the response changes with the average response;
- whether the observations are statistically independent;
- whether the response distribution under fixed covariate conditions is symmetric or skewed.

To draw inference about the parameter even based on the insufficient information as above, the likelihood function so constructed is called the Quasi-likelihood function. We concentrate mainly on the case where the observations are statistically independent and where the effects of interest can be described by a model for expected response say μ.

Some common and relaxed assumptions:

- Let the response vector: $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. The components of the response vector are independent with
  $E(\mathbf{Y}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, $\mu_i = E(Y_i)$

  And $\text{Disp}(\mathbf{Y}) = \sigma^2 \mathbf{V}(\boldsymbol{\mu})$, where
  σ^2 may be known, and
  $\mathbf{V}(\boldsymbol{\mu})$ is a matrix of known functions. Obviously
  $\mathbf{V}(\boldsymbol{\mu}) = \text{Diag}\{V_1(\boldsymbol{\mu}), \ldots, V_n(\boldsymbol{\mu})\}$ as the components of $\mathbf{Y}$ are independent $\implies$ $\text{Cov}(Y_i, Y_j) = 0 \; \forall \; i \neq j$.

- The parameter of interest β relates the dependence of $\boldsymbol{\mu}$ to covariate vector $\mathbf{x}$. Thus the symbol $\boldsymbol{\mu}(\boldsymbol{\beta})$ denotes the regression function depending on $\mathbf{x}$.
- σ^2 : constant; at least w.r.t. $\boldsymbol{\beta}$.
- $V_i(\boldsymbol{\mu}) = V_i(\mu_i)$ i.e. depends only on the ith component of $\boldsymbol{\mu}$; not over the other components.
  $\implies \mathbf{V}(\boldsymbol{\mu}) = \text{Diag}\left\{V_1(\mu_1), \ldots, V_n(\mu_n)\right\}$.

Construction of Quasi-likelihood function:

First consider iid observations $Y_1, Y_2, \ldots, Y_n$ with common mean μ and common variance $\sigma^2 V(\mu)$. Now consider the function

$$U(\mu; Y) = \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma^2 V(\mu)}$$

U has the following properties:

- $E\left[U(\mu; \mathbf{Y})\right] = 0$
- $V\left[U(\mu; \mathbf{Y})\right] = \frac{n}{\sigma^2 V(\mu)}$
- $E\left[-\frac{\partial U}{\partial \mu}\right] = V\left[U(\mu; Y)\right]$

Let us check :

$$E\left[U(\mu;\mathbf{Y})\right] = \sum_{i=1}^{n} \frac{E(Y_i)-\mu}{\sigma^2 V(\mu)} = \sum_{i=1}^{n} \frac{(\mu-\mu)}{\sigma^2 V(\mu)} = 0 \text{ (as } E(Y_i)=\mu)$$

$$\begin{aligned} V\left[U(\mu;\mathbf{Y})\right] &= \sum_{i=1}^{n} \frac{V(Y_i)}{\sigma^4 V^2(\mu)} \text{ (as } Y_i's \text{ are independent)} \\ &= \frac{n\sigma^2 V(\mu)}{\sigma^4 V^2(\mu)} \text{ (as } V(Y_i)=\sigma^2 V(\mu) \text{ under iid case)} \\ &= \frac{n}{\sigma^2 V(\mu)} \end{aligned}$$

And

$$\frac{\partial U}{\partial \mu} = \frac{V(\mu)\sum_{i=1}^{n}(-1) - V'(\mu)\sum_{i=1}^{n}(Y_i - \mu)}{\sigma^2 V^2(\mu)}$$

$\Longrightarrow$

$$E\left(-\frac{\partial U}{\partial \mu}\right) = \frac{nV(\mu) + V'(\mu)\sum_{i=1}^{n} E(Y_i) - \mu}{\sigma^2 V^2(\mu)} = \frac{nV(\mu) + V'(\mu) \times 0}{\sigma^2 V^2(\mu)}$$

$$= \frac{n}{\sigma^2 V(\mu)} = V\left[U(\mu; \mathbf{Y})\right]$$

These properties are also true for the *score function* $\frac{\partial l}{\partial \mu}$ provided

- Y_is have a distribution in the *exponential family* with form of pmf or pdf as

$$f_\theta(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

$(Member\ of\ Exponential\ Dispersion\ Family)$

where θ : Canonical parameter
$b(\theta)$: Cumulant function
ϕ : Dispersion parameter

Before realize this fact know that

$$Y \sim f_\theta(y) \implies \mu(=E(Y)) = b'(\theta) \text{ and } \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

Verify:

We know $\int f_\theta(y)dy = 1 \implies \frac{\partial}{\partial\theta}\int f_\theta(y)dy = 0$

$\iff \int \frac{\partial}{\partial\theta} f_\theta(y)dy = 0$ (*interchanging the order of integration and differentiation*)

$\iff \int \left\{\frac{\partial}{\partial\theta}\log f_\theta(y)\right\} f_\theta(y)dy = 0$

Or $E\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right) = 0 \quad \ldots \quad \ldots \quad (*)$

But here $\log f_\theta(y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$

$\implies \frac{\partial}{\partial\theta} \log f_\theta(y) = \frac{y - b'(\theta)}{a(\phi)}$

Or $E\left(\frac{\partial}{\partial\theta} \log f_\theta(Y)\right) = \frac{\mu - b'(\theta)}{a(\phi)}$ as $E(Y) = \mu$.

$(*) \implies \mu = b'(\theta)$.

Again $\frac{\partial}{\partial\theta} \int \left\{\frac{\partial}{\partial\theta} \log f_\theta(y)\right\} f_\theta(y) dy = 0$

Once again interchanging the order of integration and differentiation

$$\int \left\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(y) + \left(\frac{\partial}{\partial\theta} \log f_\theta(y) \right)^2 \right\} f_\theta(y) dy = 0$$

Continued ...

$$\implies E\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta(Y)\right) + E\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right)^2 = 0. \quad \ldots \quad \ldots \quad (**)$$

Now here $\frac{\partial^2}{\partial\theta^2}\log f_\theta(y) = -\frac{b''(\theta)}{a(\phi)}$ and $\left(\frac{\partial}{\partial\theta}\log f_\theta(y)\right)^2 = \frac{(y-b'(\theta))^2}{a^2(\phi)}$

So $(**) \implies -\frac{b''(\theta)}{a(\phi)} + \frac{\text{Var}(Y)}{a^2(\phi)} = 0$ (as $b'(\theta) = \mu$ and $E(Y-\mu)^2 = \text{Var}(Y)$)

Or $\text{Var}(Y) = b''(\theta)a(\phi) \equiv \sigma^2 V(\mu)$

[As θ expressible as a function of $\mu, \Rightarrow$ $b''(\theta) \equiv V(\mu)$; another function of μ.]

Now go back to realising the fact related to Score function $\frac{\partial}{\partial \mu} l$.

Based on a random sample $(Y_1, \ldots, Y_n)$ of size n,

the loglikelihood function $l(\theta; \mathbf{y}) = \log f_\theta(\mathbf{y}) = \dfrac{\sum_{i=1}^{n} (y_i\theta - b(\theta))}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$

So like $(*)$ and $(**)$ here also get

- $E\left(\frac{\partial}{\partial \theta} l\right) = 0 \quad \ldots \quad \ldots \quad (\otimes)$
- $E\left(\frac{\partial^2}{\partial \theta^2} l(\theta)\right) + E\left(\frac{\partial}{\partial \theta} l(\theta)\right)^2 = 0 \quad \ldots \quad \ldots \quad (\otimes\otimes)$

Also $\frac{\partial^2}{\partial\theta^2} l(\theta) = -\frac{nb''(\theta)}{a(\phi)}$

and $\frac{\partial\theta}{\partial\mu} = 1/\frac{\partial\mu}{\partial\theta} = 1/b''(\theta)$ (as $\mu = b'(\theta)$)

Now $E\left(\frac{\partial}{\partial\mu} l(\theta)\right) = \frac{\partial\theta}{\partial\mu} E\left(\frac{\partial}{\partial\theta} l(\theta)\right) = 0$ (from $(\otimes)$) ... ... $(\bullet)$

$$
\begin{aligned}
\frac{\partial^2}{\partial\mu^2} l(\theta) &= \frac{\partial}{\partial\mu}\frac{\partial}{\partial\mu} l(\theta) \\
&= \frac{\partial}{\partial\mu}\left\{\frac{\partial\theta}{\partial\mu}\frac{\partial}{\partial\theta} l(\theta)\right\} \\
&= \frac{\partial^2\theta}{\partial\mu^2}\frac{\partial}{\partial\theta} l(\theta) + \frac{\partial\theta}{\partial\mu}\frac{\partial}{\partial\mu}\frac{\partial}{\partial\theta} l(\theta) \\
&= \frac{\partial^2\theta}{\partial\mu^2}\frac{\partial}{\partial\theta} l(\theta) + \frac{\partial\theta}{\partial\mu}\frac{\partial\theta}{\partial\mu}\frac{\partial^2}{\partial\theta^2} l(\theta) \\
&= \frac{\partial^2\theta}{\partial\mu^2}\frac{\partial}{\partial\theta} l(\theta) - \frac{1}{b''(\theta)}\frac{1}{b''(\theta)}\frac{nb''(\theta)}{a(\phi)} \\
\implies E\left(-\frac{\partial^2}{\partial\mu^2} l(\theta)\right) &= \frac{n}{b''(\theta)a(\phi)} \text{ (as } E\left(\frac{\partial}{\partial\theta} l(\theta)\right) = 0 \text{ by } (\otimes))
\end{aligned}
$$

$$\implies E\left(-\frac{\partial^2}{\partial\mu^2}l(\theta)\right) = \frac{n}{\sigma^2 V(\mu)} = \frac{n}{\text{Var}(Y)} \quad \cdots \quad \cdots \quad (\bullet\bullet)$$

Now

$$\begin{aligned}\text{Var}\left(\frac{\partial}{\partial\mu}l(\theta)\right) &= E\left(\frac{\partial}{\partial\mu}l(\theta)\right)^2 - E^2\left(\frac{\partial}{\partial\mu}l(\theta)\right)\\ &= E\left(\frac{\partial}{\partial\mu}l(\theta)\right)^2 \ \left(\text{as } E\left(\frac{\partial}{\partial\mu}l(\theta)\right) = 0 \text{ by } (\bullet)\right)\end{aligned}$$

But from $(\otimes\otimes)$, $E\left(\frac{\partial}{\partial\mu}l(\theta)\right)^2 = -E\left(\frac{\partial^2}{\partial\mu^2}l(\theta)\right) = E\left(-\frac{\partial^2}{\partial\mu^2}l(\theta)\right)$.

$$\begin{aligned}\implies \text{Var}\left(\frac{\partial}{\partial\mu}l(\theta)\right) &= E\left(-\frac{\partial^2}{\partial\mu^2}l(\theta)\right)\\ &= \frac{n}{\sigma^2 V(\mu)} = \frac{n}{\text{Var}(Y)} \ (\text{using } (\bullet\bullet))\\ &\quad \cdots \quad \cdots \quad (\bullet\bullet\bullet)\end{aligned}$$

$(\bullet)$, $(\bullet\bullet)$ and $(\bullet\bullet\bullet)$ ensure the fact that $U(\mu; \mathbf{Y})$, defined earlier bears the same characteristic features as the characteristics a score function $\frac{\partial}{\partial\mu} l$ reflects in case of an exponential (-dispersion) family member.

In other words $U(\mu; \mathbf{Y})$ can play the role of a score function even when the knowledge about the probability law of the random sample observations is nil or too insufficient to construct the original score function $\frac{\partial}{\partial\mu} l$.

Quasi-Likelihood Method of Estimation

Now define

$$Q(\mu;\mathbf{Y}) = \sum_{i=1}^{n} \int_{y}^{\mu} \frac{Y_i - t}{\sigma^2 V(t)} dt$$

Observe that

$$\begin{aligned}
\frac{\partial}{\partial\mu} Q(\mu;\mathbf{Y}) &= \sum_{i=1}^{n} \frac{\partial}{\partial\mu} \int_{y}^{\mu} \frac{Y_i - t}{\sigma^2 V(t)} dt \\
&= \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma^2 V(\mu)} \\
&= U(\mu;\mathbf{Y}) = U
\end{aligned}$$

$\implies$ $Q(\mu;\mathbf{Y})$ behaves like a loglikelihood function
(*where* U *behaves like* $\frac{\partial l}{\partial\mu}$; l, *the log-likelihood function.*)

We call U the quasi-score function and Q, the quasi-likelihood function or more particularly the log quasi-likelihood function (*because it resembles to the loglikelihood function*).

These two are the elegant instrumentrs for executing the likelihood method under lack of sufficient information regarding the population probability model.

Following illustration can help for better understanding.

Illustration. Let $\sigma = 1$ and $V(\mu) = \mu(1-\mu)$, $0 < \mu < 1$. Then

$$
\begin{aligned}
Q(\mu;\mathbf{y}) &= \sum_{i=1}^{n} \int_{y_i}^{\mu} \frac{y_i - t}{t(1-t)} dt \\
&= \sum_{i=1}^{n} \left\{ y_i \int_{y_i}^{\mu} \frac{dt}{t(1-t)} - \int_{y_i}^{\mu} \frac{dt}{1-t} \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i \int_{y_i}^{\mu} \left[\frac{dt}{t} + \frac{dt}{1-t} \right] + \ln(1-t) \mid_{y_i}^{\mu} \right\} \\
&= \ln\left(\frac{\mu}{1-\mu}\right) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{1-y_i}\right) \\
&+ \quad n \ln(1-\mu) - \sum_{i=1}^{n} \ln(1-y_i)
\end{aligned}
$$

Similarly if $V(\mu) = k\mu$, $k > 1$ being known,

$$
\begin{aligned}
Q(\mu; \mathbf{y}) &= \sum_{i=1}^{n} \int_{y_i}^{\mu} \frac{y_i - t}{kt} dt \\
&= \sum_{i=1}^{n} \left\{ y_i \int_{y_i}^{\mu} \frac{dt}{kt} - \int_{y_i}^{\mu} \frac{dt}{k} \right\} \\
&= \frac{(\ln \mu + 1)}{k} \sum_{i=1}^{n} y_i - \frac{1}{k} \sum_{i=1}^{n} y_i \ln y_i - \frac{n\mu}{k}
\end{aligned}
$$

In both the examples our job is to evaluate that choice of μ that maximises $Q(\mu; \mathbf{y})$. In this context numerical method of solution can be a great help for us.

Overdispersion and Quasi-likelihood :

First we will discuss about the problem 'Overdispersion'.
According to the nature of response variable in general, we assume or fix a suitable probability distribution for it. The variance (provided it exists) of the distribution is called *Nominal Variance*.
Now, due to some additional arrangement like classification or clustering etc. on the response variable, if the variance of the distribution of response variable exceeds its nominal variance, the situation is termed as *Overdispersion*.

In reality overdispersion is not a rare or uncommon issue. In fact it will not be an exaggaration if we say that overdispersion is the real practice and having nominal variance is the exception.

Degree of overdispersion depends on the field of applications.

The problem of overdispersion encounters where the nominal variance is a function of the mean in true sense i.e. where the response distribution belongs to exponential dispersion family; a family which leads to generalised linear model of regressions under the presence of suitable covariate(s).

In reality binomial, poisson, negative binomial, gamma type response response variables mostly lead to this problem; although any type of response having distribution from the aforesaid family may experience overdispersive phenomenon.

• Below we discuss **overdispersion** problem under **binomial type response** cases:

Suppose we consider m binary responses (resulting in 0 or 1 value corresponding to 'failure' and 'success' outcome) obtained from a population having r (fixed) clusters (*Families, litters, colonies etc. are the very common examples of naturally occurring clusters*).

For symplicity, consider $k = \frac{m}{r}$ is the common cluster size (*it could vary also*).
let π_i denote the success probability in ith cluster; $0 < \pi_i < 1$; and it varies from cluster to cluster.
Let $Z_i = \#$ successes in ith cluster $\implies Z_i \sim \text{Bin}(k, \pi_i)$ independently.
Hence the total number of successes out of m Bernoulli outcomes will be

$$Y = Z_1 + Z_2 + \ldots + Z_r$$

As π_is also vary, we can assume a probability distribution (prior distribution) of π_is $\ni$ $E(\pi_i) = \pi$, say.
Now it is the most common experience that
$V(\pi_i) = c\pi(1 - \pi)$, $0 < c < 1$. **Why ?**

The most common prior of Bernoulli success probability is Beta prior. So without loss of generality suppose

$$\pi_i \overset{iid}{\sim} \text{Beta}\,(a,b)\ ,\quad a,b>0 \implies E(\pi)=\frac{a}{a+b} \text{ and } V(\pi_i)=\frac{ab}{(a+b)^2(a+b+1)}.$$

Thus $\frac{a}{a+b}=\pi \implies \frac{b}{a+b}=1-\pi$. Hence

$$V(\pi_i)=\frac{1}{(a+b+1)}\pi(1-\pi)=c\pi(1-\pi)\,,\ 0<c=\frac{1}{(a+b+1)}<1.$$

Now $E(Y)=EE(Y|\pi_i's)=E\sum_{i=1}^{r}k\pi_i=rk\pi=m\pi.$

Similarly $V(Y)=EV(Y|\pi_i's)+VE(Y|\pi_i's)=$
$E\sum_{i=1}^{r}k\pi_i(1-\pi_i)+V\sum_{i=1}^{r}k\pi_i=\sigma^2 m\pi(1-\pi)$ (on simplification) where
$\sigma^2=1+(k-1)c\geq 1$ with $=$ iff $k=1$.

Thus under $k > 1$, $V(Y) > m\pi(1-\pi)$.
But, without the knowledge of clustering we just consider $Y \sim \text{Bin}\ (m, \pi)$ which implies $V(Y) = m\pi(1-\pi)$, the nominal variance.
Hence, under $k > 1$, the consideration of binary response data from k-clustered population invites overdispersion.

Obviously the probability distribution of response Y will not have any closed form of its pmf. Thus we can successfully form here the quasi likelihood function of π and get its estimate as before.
This way overdispersive case is a relevant arena where quasi likelihood method of estimation can be successfully applied.

Similarly overdispersion can creep into the cases of Poisson type response like count data. In this context readers are suggested to go through **overdispersion** issues spread in some chapters of **Generalized Linear Models** (Second Edition) by McCullagh and Nelder for further study.